Efficient state-space modularization for planning: theory, behavioral and neural signatures

Daniel McNamee, Daniel Wolpert, Máté Lengyel
Computational and Biological Learning Lab
Department of Engineering
University of Cambridge
Cambridge CB2 1PZ, United Kingdom
{d.mcnamee|wolpert|m.lengyel}@eng.cam.ac.uk

Abstract

Even in state-spaces of modest size, planning is plagued by the "curse of dimensionality". This problem is particularly acute in human and animal cognition given the limited capacity of working memory, and the time pressures under which planning often occurs in the natural environment. Hierarchically organized modular representations have long been suggested to underlie the capacity of biological systems[1,2] to efficiently and flexibly plan in complex environments. However, the principles underlying efficient modularization remain obscure, making it difficult to identify its behavioral and neural signatures. Here, we develop a normative theory of efficient state-space representations which partitions an environment into distinct modules by minimizing the average (information theoretic) description length of planning within the environment, thereby optimally trading off the complexity of planning across and within modules. We show that such optimal representations provide a unifying account for a diverse range of hitherto unrelated phenomena at multiple levels of behavior and neural representation.

1 Introduction

In a large and complex environment, such as a city, we often need to be able to flexibly plan so that we can reach a wide variety of goal locations from different start locations. How might this problem be solved efficiently? Model-free decision making strategies[3] would either require relearning a policy, determining which actions (e.g. turn right or left) should be chosen in which state (e.g. locations in the city), each time a new start or goal location is given – a very inefficient use of experience resulting in prohibitively slow learning (but see Ref. 4). Alternatively, the state-space representation used for determining the policy can be augmented with extra dimensions representing the current goal, such that effectively multiple policies can be maintained[5], or a large "look-up table" of action sequences connecting any pair of start and goal locations can be represented – again leading to inefficient use of experience and potentially excessive representational capacity requirements.

In contrast, model-based decision-making strategies rely on the ability to simulate future trajectories in the state space and use this in order to flexibly plan in a goal-dependent manner. While such strategies are data- and (long term) memory-efficient, they are computationally expensive, especially in state-spaces for which the corresponding decision tree has a large branching factor and depth[6]. Endowing state-space representations with a hierarchical structure is an attractive approach to reducing the computational cost of model-based planning[7–11] and has long been suggested to be a cornerstone of human cognition[1]. Indeed, recent experiments in human decision-making have gleaned evidence for the use and flexible combination of "decision fragments"[12] while neuroimaging work has identified hierarchical action-value reinforcement learning in humans[13] and indicated that

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

dorsolateral prefrontal cortex is involved in the passive clustering of sequentially presented stimuli when transition probabilities obey a "community" structure[14].

Despite such a strong theoretical rationale and empirical evidence for the existence of hierarchical state-space representations, the computational principles underpinning their formation and utilization remain obscure. In particular, previous approaches proposed algorithms in which the optimal state-space decomposition was computed based on the optimal solution in the original (non-hierarchical) representation[15,16]. Thus, the resulting state-space partition was designed for a specific (optimal) environment solution rather than the dynamics of the planning algorithm itself, and also required *a priori* knowledge of the optimal solution to the planning problem (which may be difficult to obtain in general and renders the resulting hierarchy obsolete). Here, we compute a hierarchical modularization optimized for planning directly from the transition structure of the environment, without assuming any *a priori* knowledge of optimal behavior. Our approach is based on minimizing the average information theoretic description length of planning trajectories in an environment, thus explicitly optimizing representations for minimal working memory requirements. The resulting representation are hierarchically modular, such that planning can first operate at a global level across modules acquiring a high-level "rough picture" of the trajectory to the goal and, subsequently, locally within each module to "fill in the details".

The structure of the paper is as follows. We first describe the mathematical framework for optimizing modular state-space representations (Section 2), and also develop an efficient coding-based approach to neural representations of modularised state spaces (Section 2.6). We then test some of the key predictions of the theory in human behavioral and neural data (Section 3), and also describe how this framework can explain several temporal and representational characteristics of "task-bracketing" and motor chunking in rodent electrophysiology (Section 4). We end by discussing future extensions and applications of the theory (Section 5).

2 Theory

2.1 Basic definitions

In order to focus on situations which require flexible policy development based on dynamic goal requirements, we primarily consider discrete "multiple-goal" Markov decision processes (MDPs). Such an MDP, $\mathbb{M} := \{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{G}\}$, is composed of a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$ (a subset A_s of which is associated with each state $s \in \mathcal{S}$), and transition function $\mathcal{T}$ which determines the probability of transitioning to state s_j upon executing action a in state s_i, $p(s_j|s_i, a) := \mathcal{T}(s_i, a, s_j)$. A *task* (s, g) is defined by a *start state* $s \in \mathcal{S}$ and a *goal state* $g \in \mathcal{G}$ and the agent's objective is to identify a *trajectory* of *via states* $\mathbf{v}$ which gets the agent from s to g. We define a *modularization*[1] $\mathcal{M}$ of the state-space $\mathcal{S}$ to be a set of Boolean matrices $\mathcal{M} := \{M_i\}_{i=1\ldots m}$ indicating the module membership of all states $s \in \mathcal{S}$. That is, for all $s \in \mathcal{S}$, there exists $i \in 1, \ldots, m$ such that $M_i(s) = 1, \;\; M_j(s) = 0 \;\; \forall j \neq i$. We assume this to form a disjoint cover of the state-space (overlapping modular architectures will be explored in future work). We will abuse notation by using the expression $s \in M$ to indicate that a state s is a member of a module M. As our planning algorithm $\mathcal{P}$, we consider random search as a worst-case scenario although, in principle, our approach applies to any algorithm such as dynamic programming or Q-learning[3] and we expect the optimal modularization to depend on the specific algorithm utilized.

We describe and analyze planning as a Markov process. For planning, the underlying state-space is the same as that of the MDP and the transition matrix T is a marginalization over a planning policy π_{plan} (which, here, we assume is the random policy $\pi_{\text{rand}}(a|s_i) := \frac{1}{|A_{s_i}|}$)

$$T_{ij} = \sum_a \pi_{\text{plan}}(a|s_i)\, \mathcal{T}(s_i, a, s_j) \tag{1}$$

Given a modularization $\mathcal{M}$, planning at the global level is a Markov process M_{G} corresponding to a "low-resolution" representation of planning in the underlying MDP where each state corresponds

[1] This is an example of a "propositional representation"[17,18] and is analogous to state aggregation or "clustering"[19,20] in reinforcement learning which is typically accomplished via heuristic bottleneck discovery algorithms[21]. Our method is novel in that it does not require the optimal policy as an input and is founded on a normative principle.

to a "local" module M_i and the transition structure T_{G} is induced from T via marginalization and normalization[22] over the internal states of the local modules M_i.

2.2 Description length of planning

We use an information-theoretic framework[23,24] to define a measure, the (expected) *description length* (DL) of planning, which can be used to quantify the complexity of planning $\mathcal{P}$ in the induced global $L(\mathcal{P}|M_{\mathrm{G}})$ and local modules $L(\mathcal{P}|M_i)$. We will compute the DL of planning, $L(\mathcal{P})$, in a non-modularized setting and outline the extension to modularized planning DL $L(\mathcal{P}|\mathcal{M})$ (elaborating further in the supplementary material). Given a task (s, g) in an MDP, a solution $\mathbf{v}^{(n)}$ to this task is an n-state trajectory such that $\mathbf{v}_1^{(n)} = s$ and $\mathbf{v}_n^{(n)} = g$. The description length (DL) of this trajectory is $L(\mathbf{v}^{(n)}) := -\log p_{\text{plan}}(\mathbf{v}^{(n)})$. A task may admit many solutions corresponding to different trajectories over the state-space thus we define the DL of the task (s, g) to be the expectation over all trajectories which solve this task, namely

$$L(s,g) := \mathbb{E}_{\mathbf{v},n}\left[L(\mathbf{v}^{(n)})\right] = -\sum_{n=1}^{\infty}\sum_{\mathbf{v}^{(n)}} p(\mathbf{v}^{(n)}|s,g)\ \log p(\mathbf{v}^{(n)}|s,g) \tag{2}$$

This is the (s, g)-th entry of the *trajectory entropy* matrix $\mathbb{H}$ of $\mathbb{M}$. Remarkably, this can be expressed in closed form[25]:

$$[\mathbb{H}]_{sg} \quad = \quad \sum_{v\neq g}[(I - T_g)^{-1}]_{sv}\, H_v \tag{3}$$

where T is the transition matrix of the planning Markov chain (Eq. 1), T_g is a sub-matrix corresponding to the elimination of the g-th column and row, and H_v is the *local entropy* $H_v := H(T_{v\cdot})$ at state v. Finally, we define the description length $L(\mathcal{P})$ of the planning process $\mathcal{P}$ itself over all tasks (s, g)

$$L(\mathcal{P}) := \mathbb{E}_{s,g}[L(s,g)] = \sum_{(s,g)} P_s\, P_g\, L(s,g) \tag{4}$$

where P_s and P_g are priors of the start and goal states respectively which we assume to be factorizable $P_{(s,g)} = P_s\, P_g$ for clarity of exposition. In matrix notation, this can be expressed as $L(\mathcal{P}) = P_s\, \mathbb{H}\, P_g^{\mathsf{T}}$ where P_s is a row-vector of start state probabilities and P_g is a row-vector of goal state probabilities.

The planning DL, $L(\mathcal{P}|\mathcal{M})$, of a nontrivial modularization of an MDP requires (1) the computation of the DL of the global $L(\mathcal{P}|M_{\mathrm{G}})$ and the local planning processes $L(\mathcal{P}|M_i)$ for global M_{G} and local M_i modular structures respectively, and (2) the weighting of these quantities by the correct priors. See supplementary material for further details.

2.3 Minimum modularized description length of planning

Based on a modularization, planning can be first performed at the global level across modules, and then subsequently locally within the subset of modules identified by the global planning process (Fig. 1). Given a task (s, g) where s represents the *start state* and g represents the *goal state*, global search would involve finding a trajectory in M_{G} from the induced initial module (the unique M_s such that $M_s(s) = 1$) to the goal module ($M_g(g) = 1$). The result of this search will be a *global directive* across modules $M_s \to \cdots \to M_g$. Subsequently, local planning sub-tasks are solved within each module in order to "fill in the details". For each module transition $M_i \to M_j$ in M_{G}, a local search in M_i is accomplished by planning from an entrance state from the previous module, and planning until an exit state for module M_j is entered. This algorithm is illustrated in Figure 1.

By minimizing the sum of the global $L(\mathcal{P}|M_{\mathrm{G}})$ and local DLs $L(\mathcal{P}|M_i)$, we establish the optimal modularization $\mathcal{M}^*$ of a state-space for planning:

$$\mathcal{M}^* := \arg\min_{\mathcal{M}} \left[L(\mathcal{P}|\mathcal{M}) + L(\mathcal{M})\right], \text{ where } \ L(\mathcal{P}|\mathcal{M}) := L(\mathcal{P}|M_{\mathrm{G}}) + \sum_i L(\mathcal{P}|M_i) \tag{5}$$

Note that this formulation explicitly trades-off the complexity (measured as DL) of planning at the global level, $L(\mathcal{P}|M_{\mathrm{G}})$, i.e. across modules, and at the local level, $L(\mathcal{P}|M_i)$, i.e. within individual modules (Fig. 1C-D). In principle, the representational cost of the modularization itself $L(\mathcal{M})$ is also

part of the trade-off, but we do not consider it further here for two reasons. First, in the state-spaces considered in this paper, it is dwarfed by the the complexities of planning, $L(\mathcal{M}) \ll L(\mathcal{P}|\mathcal{M})$ (see the supplementary material for the mathematical characterization of $L(\mathcal{M})$). Second, it taxes long-term rather than short-term memory, which is at a premium when planning[26,27]. Importantly, although computing the DL of a modularization seems to pose significant computational challenges by requiring the enumeration of a large number of potential trajectories in the environment (across or within modules), in the supplementary material we show that it can be computed in a relatively straightforward manner (the only nontrivial operation being a matrix inversion) using the theory of finite Markov chains[22].

2.4 Planning compression

The planning DL $L(s, g)$ for a specific task (s, g) describes the expected difficulty in finding an intervening trajectory $\mathbf{v}$ for a task (s, g). For example, in a binary coding scheme where we assign binary sequences to each state, the expected length of string of random 0s and 1s corresponding to a trajectory will be shorter in a modularized compared to a non-modularized representation. Thus, we can examine the relative benefit of an optimal modularization, in the Shannon limit, by computing the ratio of trajectory description lengths in modularized and non-modularized representations of a task or environment[28]. In line with spatial cognition terminology[29], we refer to this ratio as the *compression factor* of the trajectory.

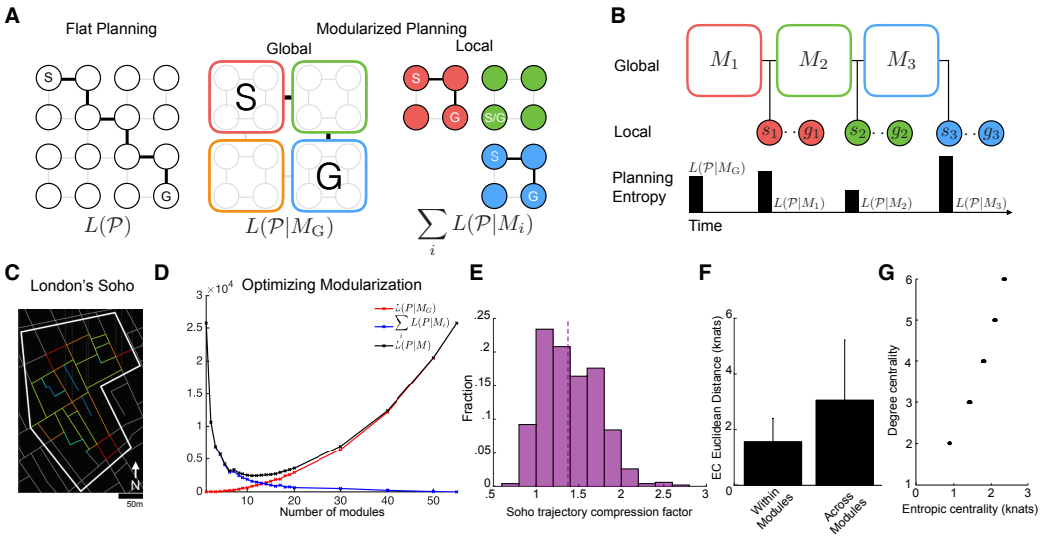

Figure 1. Modularized planning. A. Schematic exhibiting how planning, which could be highly complex using a flat state space representation (left), can be reformulated into a hierarchical planning process via a modularization (center and right). Boxes (circles or squares) show states, lines are transitions (gray: potential transitions, black: transitions considered in current plan). Once the "global directive" has been established by searching in a low-resolution representation of the environment (center), the agent can then proceed to "fill in the details" by solving a series of local planning sub-tasks (right). Formulae along the bottom show the DL of the corresponding planning processes. **B.** Given a modularization, a serial hierarchical planning process unfolds in time beginning with a global search task followed by local sub-tasks. As each global/local planning task is initiated in series, there is a phasic increase in processing which scales with planning difficulty in the upcoming module as quantified by the local DL, $L(\mathcal{P}|M_i)$. **C.** Map of London's Soho state-space, streets (lines, with colors coding degree centrality) correspond to states (courtesy of Hugo Spiers). **D.** Minimum expected planning DL of London's Soho as a function of the number of modules (minimizing over all modularizations with the given number of modules). Red: global, blue: local, black: total DL. **E.** Histogram of compression factors of 200 simulated trajectories from randomly chosen start to goal locations in London's Soho. **F.** Absolute entropic centrality (EC) differences within and across connected modules in the optimal modularization of the Soho state-space. **G.** Scatter plot of degree and entropic centralities of all states in the Soho state-space.

2.5 Entropic centrality

The computation of the planning DL (Section 2.2) makes use of the *trajectory entropy* matrix $\mathbb{H}$ of a Markov chain. Since $\mathbb{H}$ is composed of weighted sums of local entropies H_v, it suggests that we can express the contribution of a particular state v to the planning DL by summing its terms for all tasks (s, g). Thus, we define the *entropic centrality*, E_v, of a state v via

$$E_v = \sum_{s,g} D_{svg} H_v \tag{6}$$

where we have made use of the *fundamental tensor of a Markov chain* D with components $D_{svg} = \left[(I - T_g)^{-1}\right]_{sv}$. Note that task priors can easily be incorporated into this definition. The entropic centrality (EC) of a state measures its importance to tasks across the domain and its gradient can serve as a measure of "subgoalness" for the planning process $\mathcal{P}$. Indeed, we observed in simulations that one strategy used by an optimal modularization to minimize planning complexity is to "isolate" planning DL within rather than across modules, such that EC changes more across than within modules (Fig. 1F). This suggests that changes in EC serve as a good heuristic for identifying modules.

Furthermore, EC is tightly related to the graph-theoretic notion of *degree centrality* (DC). When transitions are undirected and are deterministically related to action, degree centrality $\deg(v)$ corresponds to the number of states which are accessible from a state v. In such circumstances and assuming a random policy, we have

$$E_v = \sum_{s,g} D_{svg} \frac{1}{\deg(v)} \log(\deg(v)) \tag{7}$$

The ECs and DCs of all states in a state-space reflecting the topology of London's Soho are plotted in Fig. 1G and show a strong correlation in agreement with this analysis. In Section 3.2 we test whether this tight relationship, together with the intuition developed above about changes in EC demarcating approximate module boundaries, provides a normative account of recently observed correlations between DC and human hippocampal activity during spatial navigation[30].

2.6 Efficient coding in modularized state-spaces

In addition to "compressing" the planning process, modularization also enables a neural channel to transmit information (for example, a desired state sequence) in a more efficient pattern of activity using a hierarchical entropy coding strategy[31] whereby contextual codewords signaling the entrance to and exit from a module constrain the set of states that can be transmitted to those within a module thus allowing them to be encoded with shorter description lengths according to their relative probabilities[28] (i.e. a state that forms part of many trajectory will have a shorter description length than one that does not). Assuming that neurons take advantage of these strategies in an efficient code[32], several predictions can be made with regard to the representational characteristics of neuronal populations encoding components of optimally modularized state-spaces. We suggest that the phasic neural responses (known as "start" and "stop" signals) which have been observed to encase learned behavioral sequences in a wide range of control paradigms across multiple species[33–36] serve this purpose in modularized control architectures. Our theory makes several predictions regarding the temporal dynamics and population characteristics of these start/stop codes. First, it determines a specific temporal pattern of phasic start/stop activity as an animal navigates using an optimally modularized representation of a state-space. Second, neural representations for the start signals should depend on the distribution of modules, while the stop codes should be sensitive to the distribution of components within a module. Considering the minimum average description length of each of these distribution, we can make predictions regarding how much neural resources (for example, the number of neurons) should be assigned to represent each of these start/stop variables. We verify these predictions in published neural data[36,34] in Section 4.

3 Route compression and state-space segmentation in spatial cognition

3.1 Route compression

We compared the compression afforded by optimal modularization to a recent behavioral study examining trajectory compression during mental navigation[29]. In this task, students at the University

of Toronto were asked to mentally navigate between a variety of start and goal locations on their campus and the authors computed the (inverse) ratio between the duration of this mental navigation and the typical time it would physically take to walk the same distance. Although mental navigation time was substantially smaller than physical time, it was not simply a constant fraction of it, but instead the ratio of the two (the compression factor) became higher with longer route length (Fig. 2A). In fact, while in the original study only a linear relationship between compression factor and physical route length was considered, reanalysing the data yielded a better fit by a logarithmic function ($R^2 = 0.69$ vs. 0.46).

In order to compare our theory with these data, we computed compression factors between the optimally modularized and the non-modularized version of an environment. This was because students were likely to have developed a good knowledge of the campus' spatial structure, and so we assumed they used an approximately optimal modularization for mental navigation, while the physical walking time could not make use of this modularization and was bound to the original non-modularized topology of the campus. As we did not have access to precise geographical data about the part of the U. Toronto campus that was used in the original experiment, we ran our algorithm on a part of London Soho which had been used in previous studies of human navigation[30]. Based on 200 simulated trajectories over route lengths of 1 to 10 states, we found that our compression factor showed a similar dependence on route length[2] (Fig. 2B) and again was better fit by a logarithmic versus a linear function ($R^2 = 0.82$ vs. 0.72, respectively).

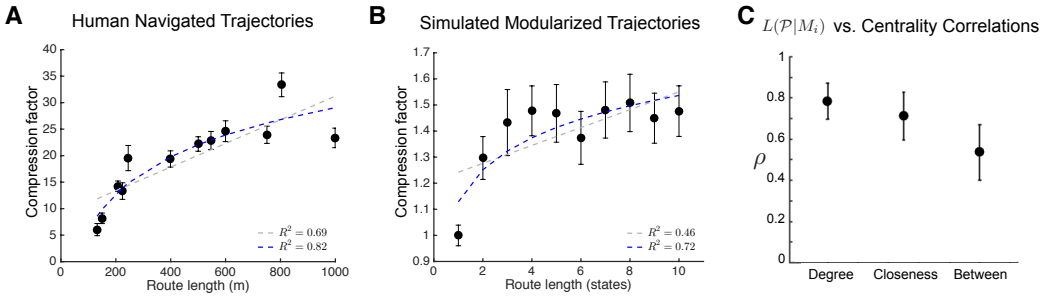

Figure 2. Modularized representations for spatial cognition. A. Compression factor as a function of route length for navigating the U. Toronto campus (reproduced from Ref. 29) with linear (grey) and logarithmic fits (blue). **B.** Compression factors for the optimal modularization in the London Soho environment. **C.** Spearman correlations between changes in local planning DL, $L(\mathcal{P}|M_i)$, and changes in different graph-theoretic measures of centrality.

3.2 Local planning entropy and degree centrality

We also modeled a task in which participants, who were trained to be familiar with the environment, navigated between randomly chosen locations in a virtual reality representation of London's Soho by pressing keys to move through the scenes[30]. Functional magnetic resonance imaging during this task showed that hippocampal activity during such self-planned (but not guided) navigation correlated most strongly with changes in a topological state "connectedness" measure known as *degree centrality* (DC, compared to other standard graph-theoretic measures of centrality such as "betweenness" and "closeness"). Although changes in DC are not directly relevant to our theory, we can show that they serve as a good proxy for a fundamental quantity in the theory, planning DL (see Eq. 7), which in turn should be reflected in neural activations.

To relate the optimal modularization, the most direct prediction of our theory, to neural signals, we made the following assumptions (see also Fig. 1B). 1. Planning (and associated neural activity) occurs upon entering a new module (as once a plan is prepared, movement across the module can be automatic without the need for further planning, until transitioning to a new module). 2. The magnitude of neural activity is related to the local planning DL, $L(\mathcal{P}|M_i)$, of the module (as the higher the entropy, the more trajectories need to be considered, likely activating more neurons with different tunings for state transitions, or state-action combinations[37], resulting in higher overall

[2]Note that the absolute scale of our compression factor is different from that found in the experiment because we did not account for the trivial compression that comes from the simple fact that it is just generally faster to move mentally than physically.

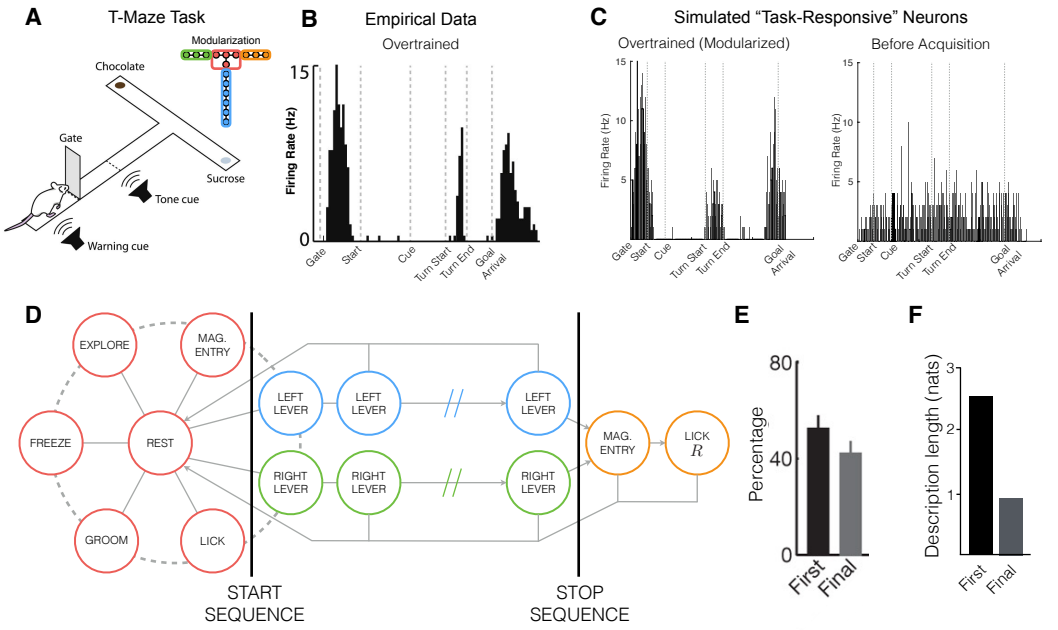

Figure 3. Neural activities encoding module boundaries. A. T-maze task in which tone determines the location of the reward (reproduced from Ref. 34). Inset: the model's optimal modularization of the discretized T-maze state-space. Note that the critical junction has been extracted to form its own module which isolates the local planning DL caused by the split in the path. **B.** Empirical data exhibiting the temporal pattern of task-bracketing in dorsolateral striatal (DLS) neurons. Prior to learning the task, ensemble activity was highly variable both spatially and temporally throughout the behavioral trajectory. Reproduced from Ref. 34. **C.** Simulated firing rates of "task-responsive" neurons after and before acquiring an optimal modularization. **D.** The optimal modularization (colored states are in the same module) of a proposed state-space for an operant conditioning task[36]. Note that the lever pressing sequences form their own modules and thus require specialized start/stop codes. **E.** Analyses of striatal neurons suggesting that a larger percentage of neurons encoded lever sequence initiations compared to terminations, and that very few encoded both. Reproduced from Ref. 36. **F.** Description lengths of start/stop codes in the optimal modularization.

activity in the population). Furthermore, as before, we also assume that participants were sufficiently familiar with Soho that they used the optimal modularization (as they were specifically trained in the experiment). Having established that under the optimal modularization entropic centrality (EC) tends to change more across than within modules (Fig. 1F), and also that EC is closely related to DC (Fig. 1G), the theory predicts that neural activity should be timed to changes in DC. Furthermore, the DLs of successive modules along a trajectory will in general be positively correlated with the differences between their DLs (due to the unavoidable "regression to the mean" effect[3]). Noting that the planning DL of a module is just the (weighted) average EC of its states (see Section 2.5), the theory thus more specifically predicts a positive correlation between neural activity (representing the DLs of modules) and changes in EC and therefore changes in DC – just as seen in experiments.

We verified these predictions numerically by quantifying the correlation of changes in each centrality measure used in the experiments with transient changes in local planning complexity as computed in the model (Fig. 2C). Across simulated trajectories, we found that changes in DC had a strong correlation with changes in local planning entropy (mean $\rho_{\text{deg}} = 0.79$) that was significantly higher ($p < 10^{-5}$, paired t-tests) than the correlation with the other centrality measures. We predict that even higher correlations with neural activity could be achieved if planning DL according to the optimal modularization, rather than DC, was used directly as a regressor in general linear models of the fMRI data.

[3]Transitioning to a module with larger/smaller DL will cause, on average, a more positive/negative DL change compared to the previous module DL.

4 Task-bracketing and start/stop signals in striatal circuits

Several studies have examined sequential action selection paradigms and identified specialized task-bracketing[33,34] and "start" and "stop" neurons that are invariant to a wide range of motivational, kinematic, and environmental variables[36,35]. Here, we show that task-bracketing and start/stop signals arise naturally from our model framework in two well-studied tasks, one involving their temporal[34] and the other their representational characteristics[36].

In the first study, as rodents learned to navigate a T-maze (Fig. 3A), neural activity in dorsolateral striatum and infralimbic cortex became increasingly crystallized into temporal patterns known as "task-brackets"[34]. For example, although neural activity was highly variable before learning; after learning the same neurons phasically fired at the start of a behavioral sequence, as the rodent turned into and out of the critical junction, and finally at the final goal position where reward was obtained. Based on the optimal modularization for the T-maze state-space (Fig. 3A inset), we examined spike trains from a simulated neurons whose firing rates scaled with local planning entropy (see supplementary material) and this showed that initially (i.e. without modularization, Fig. 3C right) the firing rate did not reflect any task-bracketing but following training (i.e. optimal modularization, Fig. 3C left) the activity exhibited clear task-bracketing driven by the initiation or completion of a local planning process. These result show a good qualitative match to the empirical data (Fig. 3B, from Ref. 34) showing that task-bracketing patterns of activity can be explained as the result of module start/stop signaling and planning according to an optimal modular decomposition of the environment.

In the second study, rodents engaged in an operant conditioning paradigm in which a sequence of eight presses on a left or right lever led to the delivery of high or low rewards[36]. After learning, recordings from nigrostriatal circuits showed that some neurons encoded the initiation, and fewer appeared to encode the termination, of these action sequences. We used our framework to compute the optimal modularization based on an approximation to the task state-space (Fig. 3D) in which the rodent could be in many natural behavioral states (red circles) prior to the start of the task. Our model found that the lever action sequences were extracted into two separate modules (blue and green circles). Given a modularization, a hierarchical entropy coding strategy uses distinct neural codewords for the initiation and termination of each module (Section 2.6). Importantly, we found that the description lengths of start codes was longer than that of stop codes (Fig. 3F). Thus, an efficient allocation of neural resources predicts more neurons encoding start than stop signals, as seen in the empirical data (Fig. 3E). Intuitively, more bits are required to encode starts than stops in this state-space due to the relatively high level of entropic centrality of the "rest" state (where many different behaviors may be initiated, red circles) compared to the final lever press state (which is only accessible from the previous Lever press state and where the rodent can only choose to enter the magazine or return to "rest"). These results show that the start and stop codes and their representational characteristics arise naturally from an efficient representation of the optimally modularized state space.

5 Discussion

We have developed the first framework in which it is possible to derive state-space modularizations that are directly optimized for the efficiency of decision making strategies and do not require prior knowledge of the optimal policy before computing the modularization. Furthermore, we have identified experimental hallmarks of the resulting modularizations, thereby unifying a range of seemingly disparate results from behavioral and neurophysiological studies within a common, principled framework. An interesting future direction would be to study how modularized policy production may be realized in neural circuits. In such cases, once a representation has been established, neural dynamics at each level of the hierarchy may be used to move along a state-space trajectory via a sequence of attractors with neural adaptation preventing backflow[38], or by using fundamentally non-normal dynamics around a single attractor state[39]. The description length that lies at the heart of the modularization we derived was based on a specific planning algorithm, random search, which may not lead to the modularization that would be optimal for other, more powerful and realistic, planning algorithms. Nevertheless, in principle, our approach is general in that it can take any planning algorithm as the component that generates description lengths, including hybrid algorithms that combine model-based and model-free techniques that likely underlie animal and human decision making[40].

References

1. Lashley K. In: Jeffress LA, editor. Cerebral Mechanisms in Behavior, New York: Wiley, pp 112–147. 1951.
2. Simon H, Newell A. Human Problem Solving. Longman Higher Education, 1971.
3. Sutton R, Barto A. Reinforcement Learning: An Introduction. MIT Press, 1998.
4. Stachenfeld K *et al. Advances in Neural Information Processing Systems*, 2014.
5. Moore AW *et al. IJCAI International Joint Conference on Artificial Intelligence* 2:1318–1321, 1999.
6. Lengyel M, Dayan P. *Advances in Neural Information Processing Systems*, 2007.
7. Dayan P, Hinton G. *Advances in Neural Information Processing Systems*, 1992.
8. Parr R, Russell S. *Advances in Neural Information Processing Systems*, 1997.
9. Sutton R *et al. Artificial Intelligence* 112:181 – 211, 1999.
10. Hauskrecht M *et al.* In: Uncertainty in Artificial Intelligence. 1998.
11. Rothkopf CA, Ballard DH. *Frontiers in Psychology* 1:1–13, 2010.
12. Huys QJM *et al. Proceedings of the National Academy of Sciences* 112:3098–3103, 2015.
13. Gershman SJ *et al. Journal of Neuroscience* 29:13524–31, 2009.
14. Schapiro AC *et al. Nature Neuroscience* 16:486–492, 2013.
15. Foster D, Dayan P. *Machine Learning* pp 325–346, 2002.
16. Solway A *et al. PLoS Computational Biology* 10:e1003779, 2014.
17. Littman ML *et al. Journal of Artificial Intelligence Research* 9:1–36, 1998.
18. Boutilier C *et al. Journal of Artificial Intelligence Research* 11:1–94, 1999.
19. Singh SP *et al. Advances in Neural Information Processing Systems*, 1995.
20. Kim KE, Dean T. *Artificial Intelligence* 147:225–251, 2003.
21. Simsek O, Barto AG. *Advances in Neural Information Processing Systems*, 2008.
22. Kemeny JG, Snell JL. Finite Markov Chains. Springer-Verlag, 1983.
23. Balasubramanian V. *Neural Computation* 9:349–368, 1996.
24. Rissanen J. Information and Complexity in Statistical Modeling. Springer, 2007.
25. Kafsi M *et al. IEEE Transactions on Information Theory* 59:5577–5583, 2013.
26. Todd M *et al. Advances in Neural Information Processing Systems*, 2008.
27. Otto AR *et al. Psychological Science* 24:751–61, 2013.
28. MacKay D. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.
29. Bonasia K *et al. Hippocampus* 26:9–12, 2016.
30. Javadi AH *et al. Nature Communications* in press, 2016.
31. Rosvall M, Bergstrom CT. *Proceedings of the National Academy of Sciences* 105:1118–1123, 2008.
32. Ganguli D, Simoncelli E. *Neural Computation* 26:2103–2134, 2014.
33. Barnes TD *et al. Nature* 437:1158–61, 2005.
34. Smith KS, Graybiel AM. *Neuron* 79:361–374, 2013.
35. Fujii N, Graybiel AM. *Science* 301:1246–1249, 2003.
36. Jin X, Costa RM. *Nature* 466:457–462, 2010.
37. Stalnaker TA *et al. Frontiers in Integrative Neuroscience* 4:12, 2010.
38. Russo E *et al. New Journal of Physics* 10, 2008.
39. Hennequin G *et al. Neuron* 82:1394–406, 2014.
40. Daw ND *et al. Nature Neuroscience* 8:1704–11, 2005.

Supplementary Material for NIPS 2016 Paper #2245 Efficient state-space modularization for planning: theory, behavioral and neural signatures

Contents

1 Modularized description length of planning

The global planning DL, $L(\mathcal{P}|M_{\mathrm{G}})$, can be easily computed after marginalizing over the internal states of each module. Defining P_{S} (P_{G}) to be the prior over start (goal) modules $P_{\mathrm{S}}(M_i) := \sum_{s\in M_i} P_s$ $\left(P_{\mathrm{G}}(M_i) := \sum_{g\in M_i} P_g\right)$, then $L(\mathcal{P}|M_{\mathrm{G}}) = P_{\mathrm{S}}\,\mathbb{H}_{\mathrm{G}}\,P_{\mathrm{G}}^{\mathsf{T}}$ where $\mathbb{H}_{\mathrm{G}}$ is the trajectory entropy across modules. In order to compute the local planning DL: $L(\mathcal{P}|M_i) := (P_{\mathrm{S}}(M_i) + P_{\mathrm{V}}(M_i) + P_{\mathrm{G}}(M_i)) \sum_{s_i,g_i\in M_i} P_{s_i}\,P_{g_i}\,L(s_i,g_i|M_i)$ we must first establish the induced priors over "sub-tasks" (s_i, g_i) within the module M_i. The probability that a state $s_i \in M_i$ serves as a sub-start state is the probability that s_i is the entrance state to M_i given a module transition into M_i at the global level. Conversely, a state $g_i \in M_i$ is a sub-goal state if it is the last transient state within M_i before a trajectory transitions out of the module M_i. These probabilities, as well as

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

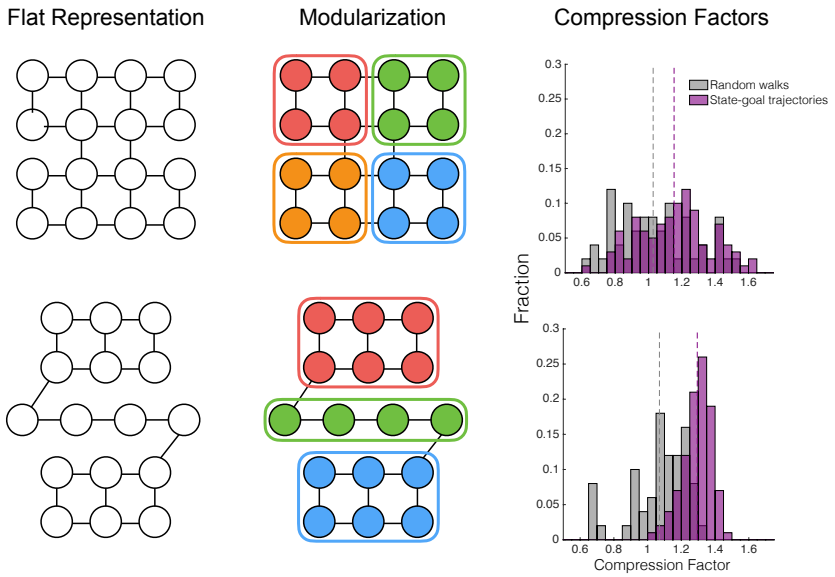

Figure 4. Two state-space modularizations exhibiting the relationship between state-space structure and modularization compression are shown. Following modularization, the first state-space is broken up into four modules corresponding to the two large "room" as well as a split in each room. Even in an a relatively homogeneous state-space (such as each of these rooms) planning complexity is minimized if the space is partitioned into equally sized modules. In the second example, the "corridor" is extracted exhibiting the partitioning of the state-space into modules with relatively constant entropic centralities. In the final column, we simulate 200 random walks and state-goal trajectories in each state-space and compute compression factors (2.4). Two types of state sequences are considered, namely a *random walk* is a random sequence of states while a *state-goal trajectory* is a sequence of states generated by a planning policy. The set of trajectories is a subset of the set of walks since the goal state cannot be repeatedly visited in a trajectory. The first example shows that the first MDP is relatively incompressible while the second exemplifies the fact that minimizing modularized DL specifically compresses solutions to problems (s, g) (i.e. trajectories) in the environment.

the probability $P_{\mathrm{V}}(M_i)$ that M_i is transiently accessed under the global directive, can be computed precisely based on the fundamental matrix of a Markov chain (see Section 2). A special case, which is computed separately, is when the start and goal states are within the same module. This contributes no additional DL at the global level but is added as a separate cost in the local planning entropy calculations.

In order to compute the optimal modularization, we currently use a brute-force algorithm, which takes around <10 mins to modularize the Soho state-space (55 states, 134 transitions). In future work, we aim to incorporate more sophisticated optimization techniques such as parallelization, greedy submodular optimization and genetic algorithms. Code is available online at https://github.com/dmcnamee.

2 Module transition probabilities

We constructed (see Section 2.5, main text) the fundamental tensor D of the global planning process

$$[D]_{SVG} := (I - T_G)^{-1}_{SV} \tag{1}$$

where S indicates a start module, V a transient module, and G a goal module. We record a useful property of the fundamental tensor. The probability that a module V is transiently accessed given a goal module G and a start module S is[1]

$$\left[D_{SVG} \times \left(\mathrm{diag}_{SV}\, D^{-1}_{SVG}\right)\right]_{SV} \tag{2}$$

This expression, weighted by the prior probabilities of tasks (S, G) gives the prior probability of a module being accessed over all tasks

$$P_V(M_V) = \sum_S [P_S D_{SVG} \times \left(\text{diag}_{SV}\, D_{SVG}^{-1}\right) P_G^T]_{SV} \tag{3}$$

We obtain the transient module transition probabilities $P(M_i \to M_j)$ by considering the global goal-module absorbing chain with fundamental matrix $N_G = (I - T_G)^{-1}$ and summing over all global tasks (S, G) weighted by the task priors P_S and P_G:

$$P(M_i \to M_j) \quad = \quad P(M_j|M_i)P(M_i) = \left[P_S D_{SVG} \times \left(\text{diag}_{SV}\, D_{SVG}^{-1}\right) P_G^T \times T|_{G^\perp,G}\right]_{ij} \tag{4}$$

2.1 Across-modules transitions and sub-start probabilities

Let us consider two connected modules, M_i and M_j, in our modularization $\mathcal{M}$ and consider the probability that an entrance state $s_j \in S_j$ is accessed from start state $s_i \in S_i$. It is known from the theory of finite Markov chains (Theorem 3.5.4 in Ref. [1]) that

$$P(s_{in} = s_j, M_j | s_{in} = s_i, M_i) = \left[N_{M_i} \times T|_{M_i,M_j}\right]_{ij} \tag{5}$$

where $T|_{M_i,M_j}$ denotes the restriction of T to the row-components corresponding to the states of M_i and the column-components of M_j. Summing over the states of $s_i \in M_i$ gives the probability $P(s_{in} = s_j, M_j|M_i)$ that $s_j \in M_j$ is an sub-start state given that the global directive has identified a transition from M_j.

2.2 Within-modules transitions and sub-goal probabilities

We assume that we have an entrance state s_a and an exit state s_b in a module M_i. The probability of s_b being an exit state from the module is the probability that it is transiently accessed before exiting to module M_j (Theorem 3.5.7 in Ref. [1])

$$P(s_{out} = s_j | s_{in} = s_i, M_i) = \left[N_{M_i} \times \text{diag}(N_{M_i})^{-1} \times T|_{M_i,M_j}\right]_{ij} \tag{6}$$

3 Alternate formulation of trajectory entropy

We use the formulation for trajectory entropy in a Markov chain established in Ref. [2]. This refines and extends a previous expression derived in Ref. [3] which we record here:

$$\mathbb{H} := K - \hat{K} + \bar{s}^{-1} L(\mathbf{v}_\infty)_\Delta \tag{7}$$

where we have

$$\begin{aligned}
K \quad &:= \quad \frac{H^* - L(\mathbf{v}_\infty)}{I - T + A^x} \\
\hat{K}_{ij} \quad &= \quad K_{jj} \ , \ \forall j \\
H^*_{ij} \quad &:= \quad H(T_{i\cdot}) \\
L(\mathbf{v}_\infty)_{ii} \quad &= \quad \bar{s}_i^{-1} H(T) \\
L(\mathbf{v}_\infty)_{ij} \quad &= \quad 0 \ , \ \forall i \neq j \\
A_{ij} \quad &= \quad \bar{s}_j
\end{aligned} \tag{8}$$

We verified numerically that these two different formulations matched in a wide range of Markov chains.

4 Representational cost of modularization

We quantify the cost of representing a modularization via the expected description length of randomly producing a particular modularization[4][5]. Such as process has two components, namely the specification of the number of modules n_M (which must be between 1 and the cardinality of the state-space

$|\mathcal{S}|$) and the assignment of each state to a module.

$$\begin{aligned} L(\mathcal{M}) &= -\log(P(\mathcal{M})) \\ &= -\log(P(\mathcal{M}|n_M)) - \log(P(n_M)) \\ &= -\log\left[\prod_{s\in\mathcal{S}} P(s \in M|n_M)\right] - \log(P(n_M)) \\ &= \sum_{s\in\mathcal{S}} \log(n_M) + \log(|\mathcal{S}|) \\ &= |S|\log(n_M) + \log(|\mathcal{S}|) \end{aligned} \tag{9}$$

5 Experimental datasets and analysis details

5.1 Spatial navigation task

Functional magnetic resonance imaging (fMRI) was used to study the brain activity of human subjects engaged in spatial navigation in London's Soho (Fig. 1C, main text). All subjects were students of University College London and therefore tended to be highly familiar with the environment. In addition, subjects were evaluated to ensure that they had prior knowledge of the environment, after completing a training process in which (1) they studied maps and photographs of the state-space locations, (2) they were given a guided tour of the area, and (3) practised the task that they would perform in the scanner[6].

On each trial, after first orienting the subjects at the start state and identifying the goal state, subjects watched first-person-view movies of travel along novel start-goal trajectories through Soho. Half of the trials required subjects to make decisions as to how to best proceed in order to complete the task. Specifically, prior to arriving at a junction in the state-space, participants indicated with a button press which subsequent direction to travel in. In control trials, subjects were instructed to press a button indicating a particular direction of travel rather than choosing themselves.

The fMRI data was analyzed with general linear models containing regressors corresponding to time series of centrality measures (betweenness, closeness, and degree) and changes thereof. The key result (see Section 3.2, main text), is that hippocampal activity was specifically sensitive to changes in degree centrality (as opposed to closeness or betweenness). Further details can be found in the publication[6] of this study.

5.2 Task-bracketing simulations

We simulated the spiking activity of Poisson neurons whose firing rate was driven by the initialization and termination of modules, and local planning entropy (within modules), in a non-modularized, and an optimally modularized version, of the T-maze state-space (Fig. 3A inset) used in Ref. 7. We assumed a baseline firing rate of 5KHz, a refractory period of 10ms, and a neural gain of 20 relating the encoded variables (start/stop, planning) to the firing rate in order to match the range of empirically observed firing rates[7]. After generating a trajectory, we resampled the time course of task variable signals to match the sampling frequency of 1000KHz. Fig. 3C (main text, modularized on the left) shows the perievent time histogram of 10 simulated trials of the ensemble activity. The median firing rate was equalized across the two conditions. We assumed that, on arrival at the goal, rodents shifted to a new behavioral module corresponding to the consummation of the reward which transiently increased planning entropy in addition to a module stop signal. Without this, there is still a clear peak at the goal arrival timepoint but with a lower average firing rate.

5.3 Operant conditioning state-space

We designed a model state-space of the operant conditioning paradigm used in Ref. 8 incorporating the fixed-ratio reward schedule relating sequences of 8 lever presses to reward delivery. In addition to the behavioral states ("lever press", "magazine entry", "lick") directly related to the action-outcome contingencies, rodents in the chamber may engage in a range of additional behaviors thus we included a range of alternative behaviors in the state-space model, namely "grooming", "resting", "freezing", and "exploring". All states connected by the dashed lines are directly accessible from one another.

For example, we assume the rodent can be in a "lick" state directly following an "explore" state without transitioning through "rest". The efficient modular decomposition displayed in Fig. 3D (main text) does not strongly depend on the structure of the state-space adjacent to the "rest" state and is mainly dependent on the natural nonuniform task distribution whereby the only rewarding goal state is "licking when reward is present" and the rodent is initialized in the "rest" state. The plotted description lengths correspond to the initialization and termination of the "lever" action sequences (as modules at the "global" level) under the stationary transition distribution.

6 Comparison with other measures of planning complexity/difficulty

Planning description length $L(\mathcal{P}|\mathcal{M})$ is a scalar measure which allows MDPs to be ranked in terms of the complexity of finding, or encoding, a solution based on a planning process $\mathcal{P}$ given a modularization $\mathcal{M}$. Note that this is distinct from the formal computational complexity theory of MDPs as a problem domain which classes them as P-complete[9]. In a set of 17 small MDPs, designed to span a variety of state-space topologies and task priors, we compared planning DL against a variety of alternative planning complexity and difficulty measures, namely (1) the expected shortest path length[10], (2) the expected path length (generated by the planning process), (3) the number of states in the MDP, (4) the number of transitions in the MDP, and (5) the average degree centrality. See Fig. 5 for scatter plots for the first four measures (see Fig. 1G, main text for a plot of entropic centrality versus degree centrality). Of these, we found that expected path length ($R^2 = 0.69$), the number of transitions ($R^2 = 0.46$), and the average degree centrality ($R^2 = 0.62$), significantly explained variability in PDL in a linear model ($p < 0.05$).

Although expected path length is significantly correlated with planning description length in our set of MDPs, it is easy to generate counter-examples to this effect. Consider an MDP consisting solely of deterministic "forward" transitions along a "corridor" of states from a start state at one end to a goal at the other (i.e. without actual choices). Here, DL agrees with intuition, assigning minimal complexity, independent of corridor length, while expected path length assigns a larger complexity, increasing with corridor length. This is exemplified by the data point in Fig. 5 with the lowest planning DL (DL equals 0, expected path length equals 6). Therefore, one can expect that state-space modularizations based on expected path length will "spend" modules on breaking up deterministic state sequences where no planning is required. Mathematically, the critical difference is the multiplication by local entropy in the planning DL measure. This sets to zero the contribution of transient states which do not contribute to overall trajectory entropy.

7 Comparing efficient modularizations and optimal behavioral hierarchies

The "optimal behavioral hierarchy"[11] (OBH) approach seeks to find the state-space decomposition which "best explains" the optimal trajectories. This objective is formalized as a bayesian model selection over the possible state-space hierarchies:

$$P(behavior|hierarchy) \propto \sum_{\pi \in \Pi} P(behavior|hierarchy, \pi) P(\pi|hierarchy) \tag{10}$$

where Π is the set of all behavioral policies which can be generated from a particular hierarchy.

This approach is distinct from that of efficient modularization (EM). First, OBH requires the optimal policy to be known before it can be applied. If used with a planning policy (such as random search) instead, as we do, it does not result in a meaningful modularization. The modularization would depend on the intrinsic stochasticity of planning via the generation of the *behavior* variable. Even if one were to optimize Eq. (10) based on the average or minimal paths of an ensemble of planning *behaviors*, such an optimized hierarchy would compress the description of such planned trajectories well but not necessarily compress the generation of them.

Second, the objectives are fundamentally different in that even if one was to use the optimal policy with EM, the modularization can be quite different from that drawn from an OBH (for examples, see Fig. 6). This is because we directly optimize for the memory requirements (see main text) whereas OBH-optimized representations would still require large capacity for maintaining the "meta-actions" of the optimal policy (in long-term memory), and for storing the resulting trajectories (in working memory). To illustrate this numerically, we established the optimal trajectories π_{opt}

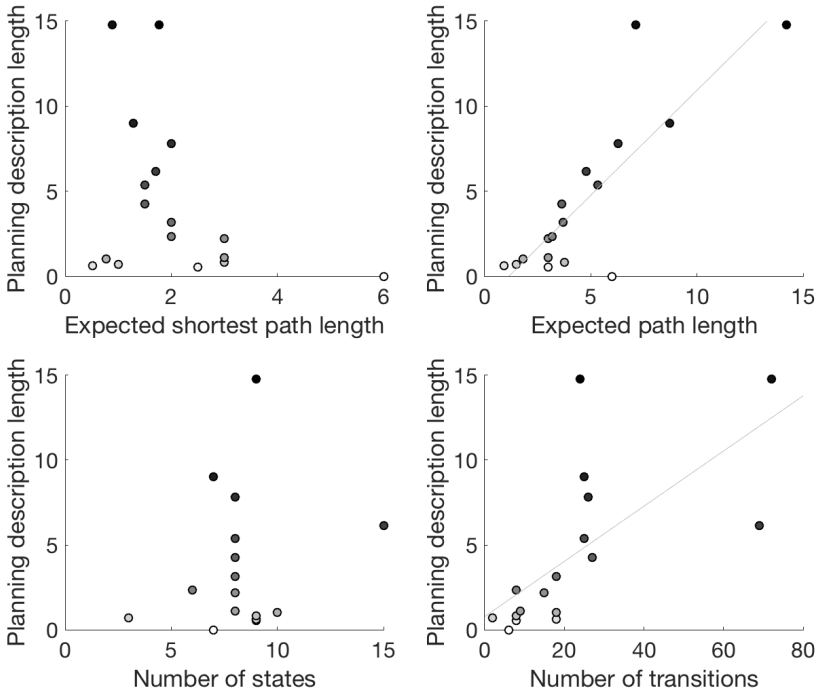

Figure 5. We computed planning description length for a variety of deterministic MDPs. For each plot, MDPs are gray-scaled in order of increasing planning DL. Significant linear relationships are indicated by a least-squares line. Planning description length is measured in nats. We describe the planning difficulty measure along the x-axes in each panel. Note that, for expected shortest path length and expected path length, the expectation of the corresponding variables under the task distribution $P(s, g)$ was used. **Expected shortest path length.** The optimal trajectories for every task (s, g) was computed and the number of states in each trajectory counted (including the goal state). **Expected path length.** The expected number of steps until arrival at the goal state under a random search planning process was computed. **Number of states.** The number of states in the MDP (independent of task prior). **Number of transitions.** The number of transitions in the MDP (independent of task prior).

(i.e. minimal paths) for all $16 \times 15 = 240$ tasks (s, g) in MDP 2 (Fig. 6), and computed the total description lengths for each trajectory: (1) L_{EM}, based on the partition defined by efficient modularization, and (2) L_{OBH}, based on the partition computed via optimal behavioral hierarchy. For all trajectories, the total description length based on EM was smaller with the average difference being $\frac{1}{240}\sum(L_{OBH}(\pi_{\text{opt}}) - L_{EM}(\pi_{\text{opt}})) = 181.69$nats. A similar analysis of trajectories generated by a random policy π_{rand} led to the same conclusion with an average difference of $\frac{1}{240}\sum(L_{OBH}(\pi_{\text{rand}}) - L_{EM}(\pi_{\text{rand}})) = 1901.33$nats. In a behavioral experiment, one could test whether the distributions of compression factors exhibited by subjects while planning in a calibrated set of MDPs and task distributions, were better fit by EM or OBH partitions.

8 Entropic centrality and state-space bottlenecks

Strongly modular decision-making environments tend to have "bottleneck" states at the interfaces between modules. From a graph-theoretic point of view, these are states which bridge between clusters of highly connected states. For planning, they serve as important "waypoints" since many trajectories must necessarily travel through them[12]. Bottlenecks are often the focus of "subgoal" discovery algorithms, based on which, temporally extended action sequences or "options" may be defined[13]. Behavioral experiments have shown[11] that human subjects can identify such bottleneck states despite only having experienced local state-state transitions and never observed the global, "bird's eye" view of the entire state-space as displayed in Fig. 7A,C.

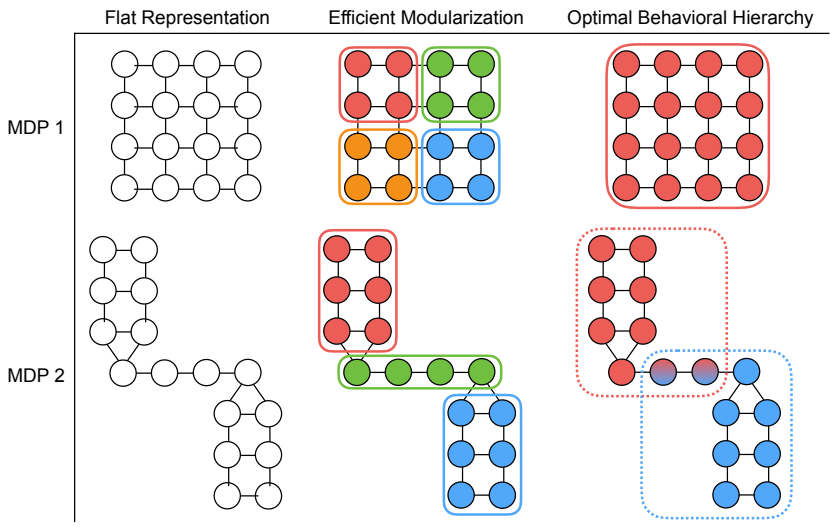

Figure 6. Maximally efficient modularizations and optimal behavioral hierarchies[11] are presented for two distinct MDPs designed to highlight differences in the corresponding partitions. For both MDPs, we assume the agent may be required to navigate between any two states with equal probability. Partitions are color-coded. **MDP 1.** This homogeneous "open space" is decomposed in the efficient modularization framework but does not contain an optimal behavioral hierarchy. The log model evidence $\log P(behavior|\text{OBH})$ (Eq. 10) in favor of the OBH hierarchy compared to the EM hierarchy is $\log P(behavior|\text{OBH}) - \log P(behavior|\text{EM}) = 168.25$. **MDP 2.** The transition structure has been altered to reflect a more modular structure (the number of states remains the same). EM extracts the "corridor" as a distinct module however the OBH has only two modules with some redundancy (the color-gradient states may be assigned, together, to either module). In this case, $\log P(behavior|\text{OBH}) - \log P(behavior|\text{EM}) = 33.08$.

It appears that efficient modularization tends to partition the environment based on changes in the entropic centrality of states (see main text). Here, we examine whether the magnitude of the entropic centrality gradient across the state-space can serve as a measure of state "bottleneckness" in an MDP using a discrete analogue of the Laplacian operator[14]. In Fig. 7A,C, we exhibit the state-space graphs of two MDPs previously used for human behavioral experiments of state bottleneck identification[11]. One can observe, from a global viewpoint, that both of these state-spaces consists of two "rooms" linked by a "corridor". Note that, in Fig. 7A, all states have the same degree centrality of three (the number of states connected to a given state). Despite this, subjects successfully[11] identified the corridor states as bottlenecks in a "bus-stop placement" task (see Ref. 11 for descriptions of the behavioral experiments).

We compute the magnitude $|\nabla E_v|$ of the entropic centrality gradient ∇E_v at state v analogously to the discrete "umbrella"[1] Laplacian operator[14] Δ based on the relation $\nabla^2 = \Delta$:

$$|\nabla E_v| = \sqrt{\sum_{n \in \mathcal{N}_1(v)} T_{nv}(E_n - E_v)^2} \tag{11}$$

where $\mathcal{N}_1(v)$ is the neighbourhood of states which are directly connected, via nonzero transitions, to state v and T is the planning transition structure of the environment (Eq. 1, main text). After computing entropic centrality gradient magnitudes at each state for each task (s, g), $|\nabla E_v|$ is the expectation of this random vector over the task prior $P(s, g)$.

In Fig. 7B,D, we scale the node sizes of the environments in Fig. 7A,C according to $|\nabla E_v|$ based on a uniform distribution of tasks (s, g) revealing how $|\nabla E_v|$ captures the degree to which a state is a

[1]This particular discrete approximation to the Laplacian operator is appropriate for our situation since the state-space has little geometric structure. If, for example, these state-spaces were embedded in a Riemannian manifold, this would induce a measure of the angle between states. Variants of Eq. 11, which incorporate more geometric structure, could be used in such a scenario.

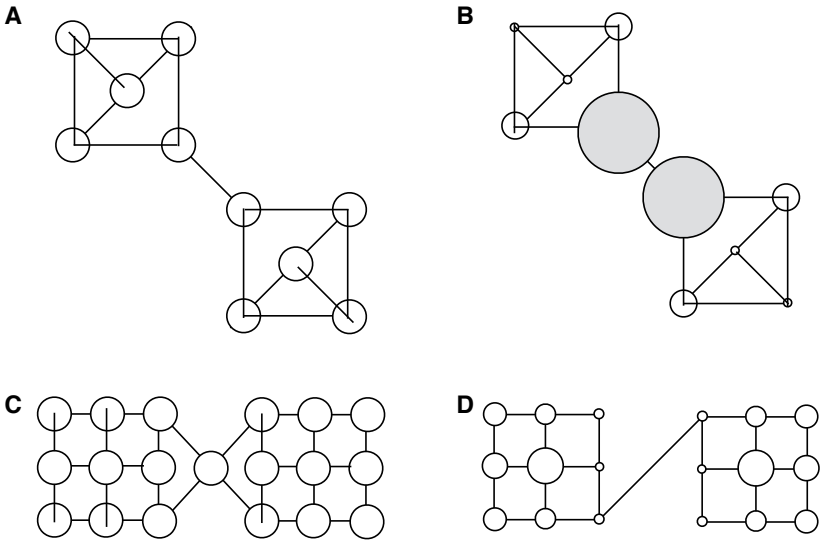

Figure 7. A. MDP used for behavioral experiments in Solway et al. (see Fig. 2C there). Importantly, this MDP has been designed such that each of the ten states has the same degree centrality (three) despite the fact that there is a "bottleneck" between the upper and lower "rooms". **B.** Node sizes are scaled according to entropic centrality gradient magnitude (Eq. 11) of the corresponding state showing that these scalar values serve as a measure of state "bottleneckness". The two state with the highest entropic centrality gradient magnitude are highlighted in grey. Interestingly, our measure ΔE also assigns the second highest value to the states which seemed to be the second most consistent choice of subjects when probed to identify bottleneck states. **C.** MDP used for behavioral experiments in Solway et al. (see Fig. 2D there). Note the clear bottleneck state between the "rooms". **D.** Node sizes are scaled according to the gradient magnitude of entropic centrality (Eq. 11). The scale is reduced compared to B in order to account for the larger number of states which globally increases entropic centrality. The state with the highest entropic centrality gradient magnitude is highlighted in grey. Our measure assigns the highest value to the bottleneck state and the second highest value to the states of high connectivity positioned at the center of the two rooms.

bottleneck in the global planning structure of the environment. This measure could aid in the discovery of subgoals, especially given that it does not require the pre-computation of the optimal policy as with previous methods[15]. Furthermore, behavioral experiments could be performed in order to test whether the apparent sensitivity of humans to state-space bottlenecks is reflective of a wider cognitive state-space representation strategy based on gradients in entropic centrality. Potentially one could implicitly infer such cognitive representations from compression factor distributions since explicitly probing subjects to reveal their perceived bottlenecks may be confounded by other considerations. For example, subjects may reasonably place a "bus-stop" specifically at the center of a long directed corridor in order to minimize the expected path length of state-goal trajectories even though this does not correspond to a bottleneck state and does not alter the complexity of planning.

References

1. Kemeny JG, Snell JL. Finite Markov Chains. Springer-Verlag, 1983.
2. Kafsi M *et al. IEEE Transactions on Information Theory* 59:5577–5583, 2013.
3. Ekroot L, Cover TM. *IEEE Transactions on Information Theory* 39:1418–1421, 1993.
4. MacKay D. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.
5. Peixoto TP. *Physical Review X* 4, 2014.
6. Javadi AH *et al. Nature Communications* in press, 2016.
7. Smith KS, Graybiel AM. *Neuron* 79:361–374, 2013.
8. Jin X, Costa RM. *Nature* 466:457–462, 2010.
9. Papadimitriou CH, Tsitsiklis JN. The Complexity of Markov Decision Processes. 1987.

10. Balaguer J *et al. Neuron* 90:893–903, 2016.
11. Solway A *et al. PLoS Computational Biology* 10:e1003779, 2014.
12. Botvinick MM *et al. Cognition* 113:262–80, 2009.
13. Barto AG, Mahadevan S. *Discrete Event Dynamic Systems* 13:41–77, 2003.
14. Wardetzky M *et al. Eurographics Symposium on Geometry Processing* pp 33–37, 2007.
15. Simsek O, Barto AG. *Advances in Neural Information Processing Systems*, 2008.