Hippocampal Spatial Model for State Space Representation in Robotic Reinforcement Learning

Angelo Arleo ANGELO.ARLEO@EPFL.CH

Wulfram Gerstner WULFRAM.GERSTNER@EPFL.CH

Lab. of Computational Neuroscience, Swiss Federal Inst. of Technology Lausanne, EPFL, CH-1015, Switzerland

Abstract

We study reinforcement learning for cognitive navigation. The state space representation is constructed by unsupervised learning during exploration. As a result of learning, a stable representation of the continuous two-dimensional manifold in the high-dimensional input space is found. The representation consists of a population of localized overlapping place fields. This state-space coding is a biologically inspired model of place fields in the Hippocampus of the rat. Place fields are learned by extracting spatio-temporal properties of the environment from visual sensory inputs. Visual ambiguities are eliminated by taking into account self-motion signals via path integration. This solves the hidden-state problem and provides a robust representation suitable for applying Q-learning in continuous space. Reward-based function approximation takes place to drive action units one synapse downstream from place cells. The teaching error models the dopaminergic reward-expectation signal from neurons in the brainstem. Several action modules share the same space representation and guide the robot to multiple targets. The experimental validation of the system is done by implementing it on a mobile Khepera robot.

1. Introduction

Problems with high-dimensional continuous state space are a critical issue in reinforcement learning (Santamaría et al., 1998). In these cases, optimizing the value function (i.e., deriving the function predicting the optimal long-term reward for a given state when selecting a specific action and following an optimal policy thenceforth) implies to learn the optimal mapping over infinitely many state-action pairs. Since exploring the whole state-action space is unfeasible, a key issue in these problems is the ability of the agent to estimate the expected value function for never experienced state-action pairs. Combining reinforcement learning with function approximation methods provides such a *generalization* property (Santamaría et al., 1998; Sutton & Barto, 1998; Lin, 1992; Rummery & Niranjan, 1994).

We study reward-based action learning to provide a robot with cognitive navigation capabilities. We ask two questions: *(i)* How can an agent learn a suitable state-space representation based on its high-dimensional and continuous sensory inputs? *(ii)* How can goal-oriented action selection be done based on the learned state-space coding?

We adopt a neuro-mimetic approach and we study spatial learning capabilities of rodents (Arleo & Gerstner, 2000). In particular, we model *(i)* the functional role of the hippocampal formation in mediating space coding in rats (O'Keefe & Nadel, 1978), *(ii)* the interaction between the hippocampus and the ventral striatum (which includes the nucleus accumbens) for mapping spatial information into actions (e.g., goal-oriented movements). Since neurons in the ventral part of the striatum are primarily activated in relation to the expectation of rewards (Schultz et al., 1992), we model such interaction through reinforcement learning.

2. State-space Representation

Neurophysiological findings suggest the spatial self-localization of rodents is supported by place-sensitive and direction-sensitive cells. *Place cells* in the rat Hippocampus provide a spatial representation in allocentric coordinates (O'Keefe & Nadel, 1978). A place cell exhibits a high firing rate only when the animal is in a specific region of the environment, which defines the *place field* of the cell. *Head-direction cells* encode the animal's allocentric heading in the azimuthal plane (Taube et al., 1990). A directional cell fires maximally only when the animal's heading is equal to the cell's *preferred direction*, regardless of the orientation of the head relative to the body, of the rat's location, or of the animal's behavior.

We present a computational model which is consistent with several neurophysiological data concerning place and head-direction cells (Arleo & Gerstner, 2000; Arleo & Gerstner, 2001). Place-coding and directional sense are provided by two coupled neural systems that interact with each other to form a unitary spatial learning system. We stress the importance of integrating allothetic (e.g., visual stimuli) and idiothetic (e.g., self-motion cues) information to establish a stable space code.

High-dimensional visual inputs (images have a resolution of 422×316 pixels) are processed by means of an unsupervised learning technique that allows the robot to detect

a convenient low-dimensional view manifold representing the visual input space. However, when relying on visual data only, the state-space representation encoded by place cells does not fulfill the Markov hypothesis (McCallum, 1996). Indeed, distinct areas of the environment may provide identical visual cues and lead to singularities in the view manifold (*sensory input aliasing*). We employ idiothetic signals (i.e., path integration) along with visual inputs in order to remove such singularities and solve the hidden-state problem. On the one hand, unreliable visual data are compensated for by means of path integration. On the other hand, reliable visual information can calibrate the path integrator system and maintain the dead-reckoning error bounded over time. Correlational learning is applied to combine visual cues and path integration over time.

After learning, the space representation consists of a population of localized overlapping place fields (Fig. 1) covering the two-dimensional workspace densely. Our place fields provide a "natural" set of basis functions in the sense that we do not have to choose parameters like width and location of the basis functions. Rather, the basis functions are created automatically by unsupervised learning. In order to interpret the information represented by the ensemble pattern of activity of our place cells, we employ *population vector coding* (Wilson & McNaughton, 1993).

3. Action Learning: Goal-oriented Navigation

Our spatial learning system provides an *incrementally learned coarse coding representation* suitable for applying reinforcement learning for continuous state spaces. Learning an action-value function over a continuous location space endows the system with spatial generalization capabilities. Since our state-space code solves the problem of ambiguous input or partially hidden states, the current state is fully known to the system and reinforcement learning can be applied in a straightforward manner.

We take a population $\mathrm{A} = \{a \,|\, 1 \leq a \leq m\}$ of *locomotor action neurons* one synapse downstream from our place cells p. Each cell a provides an allocentric directional motor command (e.g., go north). Then, the navigation problem is: How can we establish a mapping function $\mathcal{M} : \mathcal{P} \rightarrow \mathcal{A}$ from the place cell activity space $\mathcal{P}$ to the action space $\mathcal{A}$ to achieve goal-directed behavior? We apply Q-learning to approximate this function from the continuous space of physical locations to the activity space of action cells based on the agent's experience. The robot interacts with the environment, and reward-related stimuli elicit the synaptic changes of the $p \rightarrow a$ projections w_p^a to adapt the action-selection policy to the task. After training, the hippocampally-dependent activity of cells $a \in A$ provides a map to support goal-oriented navigation and obstacle avoidance.

Let r_p be the activation of a place cell p. Each state $\vec{s}$ is encoded by the ensemble activity vector $\vec{r}(\vec{s}) = (r_1(\vec{s}), \ldots, r_n(\vec{s}))$, where n is the number of place cells. The state-action value function $Q_w(\vec{s}, a)$ is of the form

$$Q_w(\vec{s}, a) = (\vec{w}^a)^T \, \vec{r}(\vec{s}) = \sum_{p=1}^{n} w_p^a \, r_p(\vec{s}) \qquad (1)$$

where $\vec{s}, a$ is the state-action pair, and $\vec{w}^a = (w_1^a, \ldots, w_n^a)$ is the adjustable parameter vector. The learning task consists of updating the vector $\vec{w}^a$ to approximate the optimal function $Q_w^*(\vec{s}, a)$. The state-value prediction error is

$$\delta_t = R_{t+1} + \gamma \, \max_a Q_t(\vec{s}_{t+1}, a) - Q_t(\vec{s}_t, a_t) \qquad (2)$$

where R_{t+1} is the immediate reward, and $0 \leq \gamma \leq 1$ is a constant discounting factor. At each time step the weight vector $\vec{w}^a$ changes according to

$$\vec{w}_{t+1}^a = \vec{w}_t^a + \alpha \delta_t \vec{e}_t \qquad (3)$$

where $0 \leq \alpha \leq 1$ is a constant learning rate parameter, and $\vec{e}_t$ is the eligibility trace vector. During learning, the exploitation-exploration trade-off is determined by an ϵ-greedy policy, with $0 \leq \epsilon \leq 1$. As a consequence, at each step t the agent might either behave greedily (*exploitation*) with probability $1 - \epsilon$, by selecting the best action a_t^* with respect to the Q-value functions, $a_t^* = \mathrm{argmax}_a Q_t(\vec{s}_t, a)$, or resort to uniform random action selection (*exploration*) with probability equal to ϵ. The update of the eligibility trace depends on whether the robot selects an exploratory or an exploiting action. Specifically, the vector $\vec{e}_t$ changes according to (we start with $\vec{e}_0 = \vec{0}$)

$$\vec{e}_t = \vec{r}(\vec{s}_t) + \begin{cases} \gamma \lambda \vec{e}_{t-1} & \text{if } \textit{exploiting} \\ 0 & \text{if } \textit{exploring} \end{cases} \qquad (4)$$

where $0 \leq \lambda \leq 1$ is a trace-decay parameter. Learning consists of a sequence of paths starting at random positions and determined by the ϵ-greedy policy. When the robot reaches the target, a new training path begins at a new random location. After learning, population vector decoding is applied to map $\mathcal{A}$ into a continuous action space $\mathcal{A}'$ by averaging the ensemble action cell activity. Fig. 2 shows an example of navigation map learned after only 5 training trials. The vector field representation of Fig. 2 has been obtained by rastering uniformly over the environment. Many of sampled locations were not visited by the robot during training which confirms the generalization capabilities of the method: the robot was able to associate appropriate goal-oriented actions to never experienced spatial states.

We assume that the anatomical counterpart of our action cells a are neurons in the nucleus accumbens (NA). NA neurons receive dopaminergic afferents related to the occurrence of unconditioned or conditioned reward-related

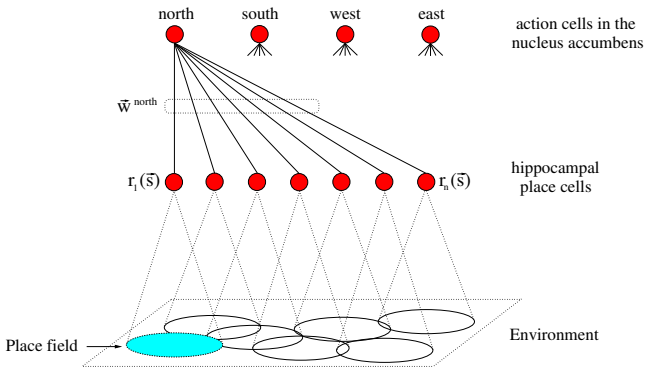

Figure 1. The overlapping place fields work as basis functions. A discrete set of actions $A = \{north, south, west, east\}$ is considered, which results in four Q-functions $Q(\vec{s}, north)$, $Q(\vec{s}, south)$, $Q(\vec{s}, west)$, $Q(\vec{s}, east)$ to be approximated.

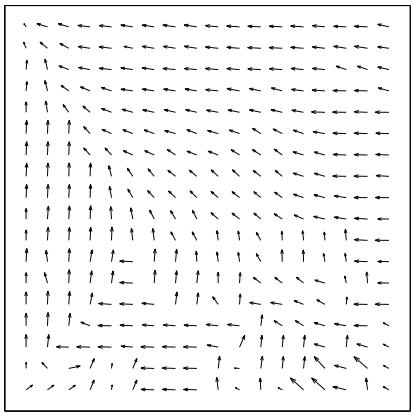

Figure 2. Vector field representation of a navigational map learned after 5 trials. The target area (about 2.5 times the area occupied by the robot) is the upper-left corner of the arena.

events (Schultz et al., 1992). The presence of dopamine-dependent plasticity in NA suggests that dopamine responses might be involved in plasticity changes yielding reward-based learning. Indeed, since the activity of dopamine neurons is a function of the unpredictability of a reward, dopaminergic activity might work as a prediction error signal (i.e., the difference $R(t) - R'(t)$ between the actual reward $R(t)$ at time t and the predicted reward $R'(t)$) suitable for learning.

4. Conclusions

Reinforcement learning takes long training time when applied directly on high-dimensional input spaces (Sutton & Barto, 1998). We have shown that by means of an appropriate state space representation, based on localized overlapping place fields, the robot can learn goal-oriented behavior after only 5 training trials. This is similar to the escape platform learning time of rats in Morris water-maze (Morris et al., 1982).

References

Arleo, A., & Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation: A model of hippocampal place cell activity. *Biological Cybernetics, Special Issue on Navigation in Biological and Artificial Systems*, *83*, 287–299.

Arleo, A., & Gerstner, W. (2001). Spatial orientation in navigating agents: Modeling head-direction cells. *Neurocomputing*, *38-40(1-4)*, 1059–1065.

Lin, L. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, *8*, 293–321.

McCallum, R. (1996). Hidden state and reinforcement learning with instance-based state identification. *IEEE Systems, Man, and Cybernetics*, *26(3)*, 464–473.

Morris, R., Garrud, P., Rawlins, J., & O'Keefe, J. (1982). Place navigation impaired in rats with hippocampal lesions. *Nature*, *297*, 681–683.

O'Keefe, J., & Nadel, L. (1978). *The Hippocampus as a cognitive map*. Oxford: Clarendon Press.

Rummery, G., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (Technical Report CUED/F-INFENG/TR 166). Engineering Department, Cambridge University.

Santamaría, J., Sutton, R., & Ram, A. (1998). Experiments with reinforcement learning in problems with continuous state and action spaces. *Adaptive Behavior*, *6(2)*, 163–218.

Schultz, W., Apicella, P., Scarnati, E., & Ljungberg, T. (1992). Neuronal activity in monkey ventral striatum related to the expectation of reward. *Journal of Neuroscience*, *12*, 4595–4610.

Sutton, R., & Barto, A. (1998). *Reinforcement learning, an introduction*. Cambridge, Massachusetts: MIT Press-Bradford Books.

Taube, J., Muller, R., & Ranck, J. (1990). Head direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *Journal of Neuroscience*, *10*, 420–435.

Wilson, M., & McNaughton, B. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, *261*, 1055–1058.