Hippocampal formation breaks combinatorial explosion for reinforcement learning: A conjecture

András Lőrincz
Department of Information Systems, Eötvös Loránd University
Pázmány Péter sétány 1/C
Budapest, Hungary H-1117

Abstract

There is surmounting evidence that reinforcement learning (RL) is a good model for the dopamine system of the brain and the prefrontal cortex. RL is also promising from the algorithmic point of view, because recent factored RL algorithms have favorable convergence and scaling properties and can counteract the curse of dimensionality problem, the major obstacle of practical applications of RL methods. The learning task then separates (i) to the search and the encoding of the factors, such as direction, position, form, and so on, and (ii) to the optimization of decision making, i.e., RL, by using these factors. We conjecture that the main task of the hippocampal formation is to support goal oriented behavior by separating factors for RL and then encoding the factors into neocortical areas. The necessary mathematical framework is sketched that includes convergent factored RL model and the method for finding independent factors. We present the factor forming comparator network model of the hippocampal formation.

Introduction

RL has been strongly motivated by psychology and neuroscience (Schultz 2002; Daw, Niv, and Dayan 2005; Delgado 2007). In addition, RL has successfully reached human level in different games (Tesauro 1995; Conitzer and Sandholm 2007; Szita and Lőrincz 2007).In this paper we assume that reinforcement learning (RL) is the cornerstone in modeling the brain.

The core problem of RL is that learning in simple environments needs information about a number of variables, such as the number of objects present, their shape or color, their distances, speeds, directions, other agents, including relatives, friends, and enemies, and so on. The state space grows with the number of variables in the exponent. Another source of combinatorial explosion is that the agent needs to maintain a history of sensory information for learning and the temporal depth multiplies the exponent. That is, advanced RL algorithms featuring near optimal polynomial time convergence in the number of states (Kearns and Singh 1998) are still troubled, because the state space can be exponentially large.

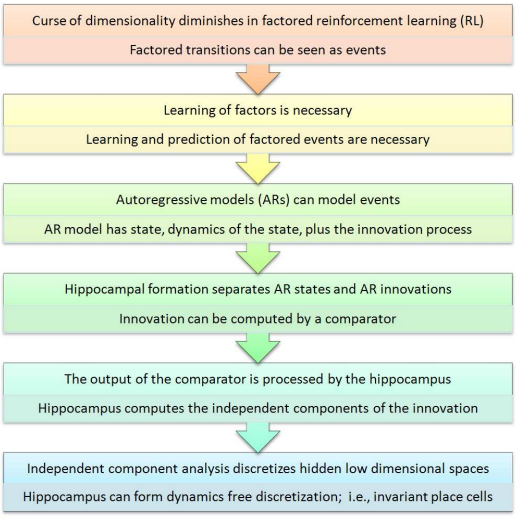


Figure 1: Flowchart about the structure of the arguments

Within the framework of RL, factored description can decrease the exponent of the state space (Boutilier, Dearden, and Goldszmidt 1995; 2000; Koller and Parr 2000; Guestrin et al. 2003), because only the relevant variables of the state need to be considered. However, the number of equations may still explode.

We describe a novel function approximation approach to factored RL that converges in polynomial time in spite of the above mentioned problems. Factored RL needs the discovery of factors, too. We argue that autoregressive (AR) independent process analysis (IPA) is an appropriate candidate for this search, because these processes can form spatio-temporal chunks that eliminate the explosion coming from the temporal depth problem of the Markov condition of RL. We map the AR-IPA algorithm to the loop of the entorhinal cortex (EC) and the hippocampus (HC). Numerical experiments show intriguing agreements with neuronal properties measured in the EC-HC loop of rats providing implicit support to the mapping. The paper ends with a short conclusion. Figure 1 depicts the flow of thoughts.

Factored reinforcement learning

A Markov Decision Process (MDP), the core of RL models, is characterized by a sixtuple $(\mathbf{X}, A, R, P, \mathbf{x}_s, \gamma)$, where $\mathbf{X}$ is a finite set of states; A is a finite set of possible actions; $R : \mathbf{X} \times A \to \mathbb{R}$ is the reward function of the agent, so that $R(\mathbf{x}, a)$ is the reward of the agent after choosing action a in state $\mathbf{x}$; $P : \mathbf{X} \times A \times \mathbf{X} \to [0,1]$ is the transition function so that $P(\mathbf{y} \mid \mathbf{x}, a)$ is the probability that the agent arrives at state $\mathbf{y}$, given that she started from $\mathbf{x}$ upon executing action a; $\mathbf{x}_s \in \mathbf{X}$ is the starting state of the agent; and finally, $\gamma \in [0,1)$ is the discount rate on future rewards.

A policy of the agent is a mapping $\pi : \mathbf{X} \times A \to [0,1]$ so that $\pi(\mathbf{x}, a)$ tells the probability that the agent chooses action a in state $\mathbf{x}$. For any $\mathbf{x}_0 \in \mathbf{X}$, the policy of the agent and the parameters of the MDP determine a stochastic process experienced by the agent through the instantiation

$$\mathbf{x}_0, a_0, r_0, \mathbf{x}_1, a_1, r_1, \ldots, \mathbf{x}_t, a_t, r_t, \ldots$$

The goal is to find a policy that maximizes the expected value of the discounted total reward. Let the *state value* function of policy π be

$$V^\pi(\mathbf{x}) := E\Big(\sum_{t=0}^{\infty} \gamma^t r_t \;\Big|\; \mathbf{x} = \mathbf{x}_0\Big)$$

and let the optimal value function be

$$V^*(\mathbf{x}) := \max_\pi V^\pi(\mathbf{x})$$

for each $\mathbf{x} \in \mathbf{X}$. If V^* is known, it is easy to find an optimal policy π^*, for which $V^{\pi^*} \equiv V^*$. Because history does not modify transition probability distribution $P(\mathbf{y}|\mathbf{x}, a)$ at any time instant, value functions satisfy the famous Bellman equation

$$V^*(\mathbf{x}) = \max_a \sum_{\mathbf{y}} P(\mathbf{y} \mid \mathbf{x}, a)\Big(R(\mathbf{x}, a) + \gamma V^*(\mathbf{y})\Big). \qquad (1)$$

Most algorithms that solve MDPs build upon some version of the Bellman equations. For example, value can be computed by an iterative solution to the Bellman equation starting from an arbitrary value function $V_0 : \mathbf{X} \to \mathbb{R}$, and in iteration t performing the update

$$V_{t+1}(\mathbf{x}) := \max_a \sum_{\mathbf{y} \in \mathbf{X}} P(\mathbf{y} \mid \mathbf{x}, a)\Big(R(\mathbf{x}, a) + \gamma V_t(\mathbf{y})\Big) \qquad (2)$$

for all $\mathbf{x} \in \mathbf{X}$.

Relevant results

Due to the space limitations, detailed derivations are omitted, but see the supplementary material (Szita and Lőrincz 2008). Here follows the list of the relevant facts, results, and tasks to be accomplished.

1. MDP concepts can be generalized to *factored Markov decision processes* (Boutilier, Dearden, and Goldszmidt 1995), where $\mathbf{X}$ is the Cartesian product of m smaller state spaces (corresponding to individual variables):

$$\mathbf{X} = X_1 \times X_2 \times \ldots \times X_m.$$

where the size of X_i is $|X_1| = n_i$, $i = 1, 2 \ldots m$. The size of the full state space is $N = |\mathbf{X}| = \prod_{i=1}^m n_i$. The number of variables is limited by the number of factors necessary to characterize the transition function.

2. *State values*, or similarly, *state-action values* can be approximated by means of basis functions. In particular, approximate value iteration with linear function approximations have attractive properties. Convergent forms can be found in (Szita and Lőrincz 2008) and it is also shown that these convergent forms can be carried over to case of factored value iteration.

3. Although the number of equations is still exponentially large for the factored case, sampling methods can be exploited (Szita and Lőrincz 2008). Similarly to the results of (Drineas, Mahoney, and Muthukrishnan 2006), the result of sampling is that the full algorithm remains polynomial.

4. Factored RL algorithms assume that factors are already known, but in the general case, factors should be searched for and learned. We assume that factors correspond to AR processes. AR processes do appear in RL if we combine policy and dynamics into a predictive matrix. Furthermore, optimization can take advantage of generating virtual experiences in this case (Sutton et al. 2008).

5. The circle is not closed yet, there are missing links. For example, (i) AR processes may form the basis of RL only if decisions concern the launching or the termination of the processes. These are desired properties concerning ongoing events, which may succeed or may not. The so called event-based description (Szita, Takács, and Lőrincz 2003) seems appropriate for this formalization. Furthermore, (ii) different AR processes may coexist and decision making may concern low-complexity combinations (i.e., a small number of the processes) at a time. The appropriate formalism, low-complexity rule based policy iteration, have been developed (Szita and Lőrincz 2007). It should be extended to low-complexity AR-based policy iteration.

6. The envisioned learning sequence uncovers other missing pieces. Firstly, the factors should be uncovered and the relevant factors should be selected. The values of turning factors on or off should be estimated. This estimation calls for the extension of convergence proof of the factored value iteration method (Szita and Lőrincz 2008) to that of factored temporal difference (TD) learning, which has shown attractive properties in numerical studies (Gyenes, Bontovics, and Lőrincz 2008). Transition probabilities need to be estimated in order to take advantage of linear Dyna-like algorithms (Sutton et al. 2008) in virtual experiences. This is a necessary prerequisite for goal oriented multi-agent scenarios, if novel information concerning the rewards is communicated by (at least one of) the agents.

Autoregressive processes

We are to find independent AR processes and their driving sources. Independent component analysis (ICA) and independent subspace analysis (ISA) are the key concepts for us. We start with a note about ICA, turn to ISA and finally to AR processes.

Note about ICA. ICA has a peculiar but sensible property for large number of distributed sensors when the underlying space is low dimensional. In this case, ICA minimizes mutual information by performing an apparent discretization of the low dimensional space (Lőrincz et al. 2001; Takács and Lőrincz 2006). This property is important for factored RL.

Independent Subspace Analysis. ISA (Cardoso 1998) is a generalization of ICA. ISA assumes that certain sources depend on each other, but the dependent groups of sources are independent of each other, i.e., the independent groups are multidimensional. The ISA task has been subject of extensive research (see, e.g., (Szabó, Póczos, and Lőrincz 2007) and the references therein). ISA is important in our modeling for the following reasons:

7. For groups of sources that are statistically independent of other ones, characterization of the sources can be restricted to the subspace of the variables making considerable savings in the exponent of the state space.

8. Cardoso's observation (Cardoso 1998) that components found by ICA are entailed by single independent subspaces has received mathematical support for a limited set of distributions (Szabó, Póczos, and Lőrincz 2007). This observation enables effective non-combinatorial search for independent subspaces, without a priori information about the number of subspaces and their dimensions. Thus, the non-combinatorial algorithm provides combinatorial gains (Póczos et al. 2007).

ISA can be used to reveal the independent source groups of AR processes, so we can influence the processes by interacting with the sources. Let $\mathbf{e}(t)$ denote the concatenated vector of the source components $\mathbf{e}^m(t) \in \mathbb{R}^{d_e^m}$. The total dimension of the components is $D_e = \sum_{m=1}^{M} d_e^m$. We assume that for a given m, $\mathbf{e}^m(t)$ is i.i.d. in time t, and sources $\mathbf{e}^m$ are jointly independent, i.e., $I(\mathbf{e}^1,\ldots,\mathbf{e}^M) = 0$, where $I(.)$ denotes the mutual information of the arguments. We shall see that the decomposition comes for non-combinatorial costs. First we characterize the ISA task and then we extend it to the AR identifications task.

The ISA task can be formalized as follows:

$$\mathbf{x}(t) = \mathbf{A}\mathbf{e}(t), \text{ where } \mathbf{e}(t) = \left[\mathbf{e}^1(t);\ldots;\mathbf{e}^M(t)\right] \in \mathbb{R}^{D_e} \qquad (3)$$

The dimension of the observation $\mathbf{x}$ is D_x. Assume that $D_x > D_e$, and $\mathbf{A} \in \mathbb{R}^{D_x \times D_e}$ has rank D_e. Then, one may assume without any loss of generality that both the observed ($\mathbf{x}$) and the hidden ($\mathbf{e}$) signals are white, i.e., decorrelated and normalized.

We are to uncover the independent subspaces. Our task is to find orthogonal matrix $\mathbf{W} \in \mathbb{R}^{D_e \times D_x}$ such that $\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t)$, $\mathbf{y}(t) = \left[\mathbf{y}^1(t);\ldots;\mathbf{y}^M(t)\right]$, $\mathbf{y}^m = [y_1^m;\ldots;y_{d_e^m}^m] \in \mathbb{R}^{d_e^m}$, $(m = 1,\ldots,M)$ with the condition that components $\mathbf{y}^m$ are independent. Here, y_i^m denotes the i^{th} coordinate of the m^{th} estimated subspace. This task can be solved by means of a cost function that aims to minimize the mutual information between components:

$$J_1(\mathbf{W}) \doteq I(\mathbf{y}^1,\ldots,\mathbf{y}^M). \qquad (4)$$

One can rewrite $J_1(\mathbf{W})$ as follows:

$$J_2(\mathbf{W}) \doteq I(y_1^1,\ldots,y_{d_e^M}^M) - \sum_{m=1}^{M} I(y_1^m,\ldots,y_{d_e^m}^m). \qquad (5)$$

The first term of the r.h.s. is an ICA cost function; it aims to minimize mutual information for all coordinates. The other term is a kind of *anti-ICA* term; it maximizes mutual information within the subspaces. One may try to apply a heuristics and to optimize (5) in the following order: (i) Start by any 'infomax'-based ICA algorithm and minimize the first term of the r.h.s. in (5). (ii) Apply only permutations on the coordinates to optimize the second term. Surprisingly, this heuristics leads to the global minimum of (4) in many cases. In other words, ICA that minimizes the first term of the r.h.s. of (5) solves the ISA task as well, apart from grouping the coordinates into subspaces. This feature was first observed by Cardoso (Cardoso 1998). To what extent this heuristic works is still an open issue. Nonetheless, we consider it as a '*Separation Theorem*', because for elliptically symmetric sources and for some other distribution types one can prove that it is rigorously true (Szabó, Póczos, and Lőrincz 2007).

Another issue concerns the computation of the second term of the r.h.s. of (5). For subspaces $\mathbf{e}^m$ of known dimensions d_e^m, multi-dimensional entropy estimations can be applied, but these are computationally expensive. Non-combinatorial estimation of the number of components and their respective dimensions have been developed recently (Póczos et al. 2007).

Identification of independently driven AR processes The generalization of the ISA task to the AR identification task in its simplest form is this

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t), \qquad (6)$$

$$\mathbf{s}(t+1) = \mathbf{F}\mathbf{s}(t) + \mathbf{e}(t), \qquad (7)$$

and it can be extended to higher order AR processes, moving average processes, integrated processes and the combinations (Szabó, Póczos, and Lőrincz 2007). The identification algorithms can be put into Hebbian (that is neurally plausible) forms (Lőrincz, Kiszlinger, and Szirtes 2008). These Hebbian forms pose constraints that can be used to explain several intriguing properties of the hippocampal formation (HF). We describe the HF and the related model below.

Model of the hippocampal formation

The relevance of the hippocampal formation. The HF seems quite similar in mammals and it is generally considered to play crucial role in the formation of declarative memory; the memory about the representations of facts, events or rules. Interestingly, after lesion, remote memories remain mostly intact but new memories about events that occurred after the lesion can not be formed (Squire 1992; Scoville and Milner 1957). In other words, the HF is required for the acquisition (learning or creating representations), but then it carries over the information to other regions so the information can be preserved even without the HF. It is also relevant that other forms of learning, such as learning of procedures, or category learning remain intact after hippocampal lesion.

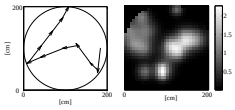


Figure 2: (a): Circular maze, diameter: 2m, with a short sample trajectory. Step size varies between 11 and 55 cm. (b): Sample input to the loop in the form of an activity map within the maze. Activity map is shown in arbitrary units.

We consider events and facts as spatio-temporal chunks that can be combined under certain conditions. Factors of the events may include for example, position, direction and speed of the animal, which together provide a crude description of its dynamical state. Many other factors can be envisioned, e.g., objects if they are present, their color and shape, if the animal is hungry or thirsty or both and so on. The rat's HF may contain place cells (mostly) invariant to rotation, It may also contain head direction cells, which have certain, but typically weak spatial modulations, and so on. It looks that these cell ensembles, which are located at different regions of the HF encode factors, whereas the cells themselves form a crude and soft discretization of the respective factors.

ICA model of the HF. Inspired by early ideas (Attneave 1954; Barlow 1961) ICA has been suggested to take place in the HF (Lőrincz 1998). The idea has been improved and extended over the years (Lőrincz and Buzsáki 2000; Chrobak, Lőrincz, and Buzsáki 2000; Lőrincz, Szatmáry, and Szirtes 2002; Franzius, Sprekeler, and Wiskott 2007) and the resulting model seems powerful enough to predict intriguing features of this region, e.g., the independence of neuronal firing in certain areas (Redish et al. 2001), long and tunable delays in other places (Henze, Wittner, and Buzsáki 2002) and different functional roles for different pathways of the loop (Kloosterman, van Haeften, and da Silva 2004).

The model is flexible and can explain novel findings about certain properties of the rat HF (McNaughton et al. 2006)). We designed computer simulations to highlight the computational roles and the related representations of the different parts of the HF.

Numerical simulations. In our simulations[1], the 'rat' is 'walking' in a circular arena and observes distributed mixed scent patches (Fig. 2).

Information was cumulated on the trajectories and was analyzed by AR-IPA tools. We found that the ICA transform of temporally integrated information resemble to the response of head direction cells. The fitted AR process produces artificial neurons that represent a low-quality grid with conjunctive position and direction information. The neural representation of the innovation of the AR process forms a reasonably improved hexagonal grid and the ICA transform of this grid brings about place cells. Because inputs are direction invariant scents, they have no predictive power without temporal integration. The estimated AR matrix F is zero on this input, and this process corresponds to an innovation process. Results for this case are shown in Fig. 3

[1]For a detailed description of the simulations see the online supplementary material (Lőrincz, Kiszlinger, and Szirtes 2008).

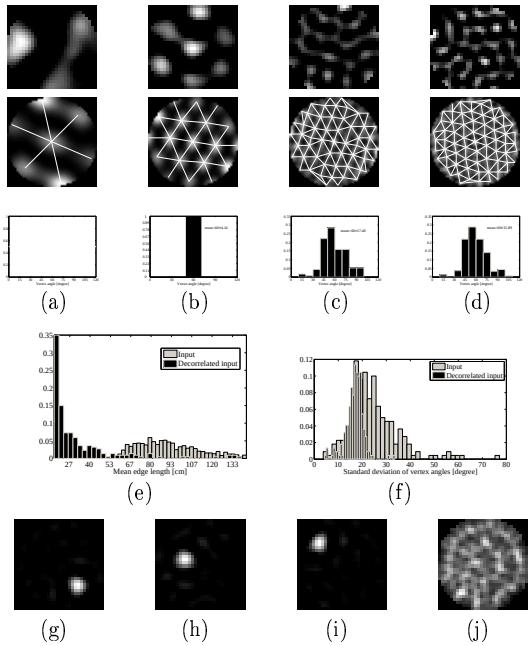
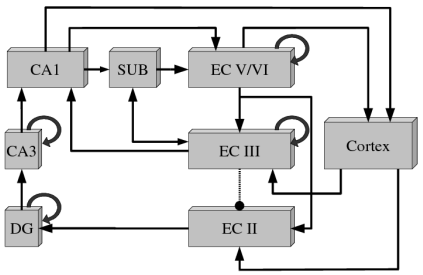


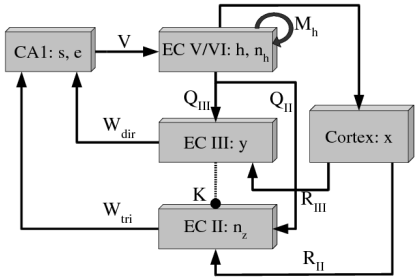
Figure 3: Position dependent input. (a-d): each column shows the output of different *decorrelating* units (PCA). First row: half-wave rectified and scaled activity maps (0: black, 1: white). Second row: 2D autocorrelation function of the activity maps and the fitted grids. Third row: vertex angle histogram for the fitted grids. (e-f): cumulative statistics over all grids. (e): histogram of the mean edge length of the grids for the input set and the PCA units. (f): histogram of the standard deviation of the vertex angles for the input set and the PCA units. (g-i): sign corrected activity maps of three *separating* (ICA) units. Response is localized. (j): superimposed map of all ICA units demonstrating that the localized units cover the full maze. For more details, see, (Lőrincz, Kiszlinger, and Szirtes 2008)

The mapping of the AR-IPA process to the HF is sketched in Fig. 4, but many of the details are omitted, including the peculiar role of the dentate gyrus, the Hebbian learning rules for the hidden process that require two-phase operation in the comparator loop (Lőrincz and Szabó 2007), an important property of the HF. For these and other details, see the supplementary material (Lőrincz, Kiszlinger, and Szirtes 2008).

In sum, results point to the proper direction, although the precision of the hexagonal grid found in the numerical studies is not yet satisfactory. It seems hard to achieve the required precision without motion related internal measures. In addition, the brain somehow takes motion related information apart, e.g., information, which is invariant to straight motion and information, which shows rotation invariance. These components are physically separated; they are encoded in different areas. It is unclear how to separate these factors into different brain regions. In our conjecture, state of the AR process is characterized by direction and weak spatial information, whereas innovation represents a differ-



(a)



(b)

Figure 4: Hippocampal formation and the model: the factor forming structure. *Factored RL estimation* occurs in a separate architecture, not shown here. For the prototype of the RL architecture see, (Szita, Gyenes, and Lőrincz 2006).
(a): Diagram of the main connections of hippocampus (HC) and its environment. Arrows denote excitatory connections and solid circles denote mostly inhibitory connections. HC entails the dentate gyrus (DG), the CA1 and the CA3 subfields. Respective layers of the entorhinal cortex (EC) are denoted by roman letters. The CA1 subfield approximates the position and holds its discretization, the *place cells*. The subiculum approximates *head directions* and holds the corresponding discretization. (b): Connections and different representations playing a role in the numerical simulations:
x: signal from cortex,
y: whitened (decorrelated and normalized) input at EC III,
$\mathbf{n_z}$: whitened novelty (i.e., innovation) of the input at EC II,
h: hidden model at EC deep layers,
$\mathbf{n_h}$: innovation of the hidden model at EC deep layers,
s: ICA output – the multivariate AR process – at CA1 during positive theta phase,
e: ICA output – the innovation – at CA1 during negative theta phase,
R_{II} and R_{III}: postrhinal to EC II and postrhinal to EC III efferents, respectively,
Q_{II} and Q_{III}: EC deep layers to EC II and EC III connections, respectively,
K inhibitory feedback from EC III to EC II.
V: CA1 to EC deep layer efferents,
M_h: recurrent collaterals at the deep layers of the EC,
W_{tri} and W_{dir}: tri-synaptic and direct connections between EC superficial layers and the CA1 subfield, respectively.

ent factor. These factors are discretized by the ICA algorithm. It is probable that actions, that is, control-like information play a role in the forming of the representations. Also, the complementarity of the proximal and distal afferents of the subiculum and the CA1 subfield of the hippocampus may be necessary for the separation of motion related factors *and* performing ICA on these separated factors. These are open problems of the model.

Conclusions

We started from the necessity of RL, showed that attractive convergent polynomial learning algorithms exists provided that approximate factored representation can be found. We have argued that factors are spatial-temporal chunks and factored representation concerns the combination of such chunks. In this conjecture, decision making is concerned with the transitions among sets of factors, alike to the method described in (Szita and Lőrincz 2007).

Then the trick is to learn to separate the chunks that can be combined later. It has been proposed that the EC-HC loop represents an AR model with the innovation held in EC layer II (Lőrincz and Buzsáki 2000; Chrobak, Lőrincz, and Buzsáki 2000). Recent advances on AR-IPA theory (Szabó, Póczos, and Lőrincz 2007; Lőrincz and Szabó 2007) underline the option that the HC could perform ICA. Here, we suggested that the state and the innovation of the AR process may form (some of the) factors for RL and that the factors are discretized by ICA.

Numerical simulations indicate that both hexagonal grids and place cells can be developed under suitable conditions in the model. The same simulations also point to the need to extend the information available for ICA by control variables. The challenge for the model is to include control values such that the precision of the hexagonal grid improves and show the intriguing property that grids faithfully follow the distortion of the (familiar) environment (Barry et al. 2007) of grid cells when the arena is distorted.

Acknowledgement

This research has been supported by the EC FET Grant FP6-502386 and EC NEST Grant FP6-043261 and Air Force Grant FA8655-07-1-3077. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the European Commission, Air Force Office of Scientific Research, Air Force Research Laboratory.

References

Attneave, F. 1954. Some informational aspects of visual perception. *Psychological Review* 61:183–193.

Barlow, H. B. 1961. *Sensory Communication*. MIT Press, Cambridge, MA. 217–234.

Barry, C.; Hayman, R.; Burgess, N.; and Jeffery, K. J. 2007. Experience-dependent rescaling of entorhinal grids. *Nature Neuroscience* 10(6):682–684.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 1995. Exploiting structure in policy construction. In *Proc. of the 14th Int. Joint Conf. on Artif. Intell.*, 1104–1111.

Boutilier, C.; Dearden, R.; and Goldszmidt, M. 2000. Stochastic dynamic programming with factored representations. *Artif. Intell.* 121:49–107.

Cardoso, J. 1998. Multidimensional independent component analysis. In *Proc. of ICASSP*, volume 4, 1941–1944.

Chrobak, J. J.; Lőrincz, A.; and Buzsáki, G. 2000. Physiological patterns in the hippocampo-entorhinal cortex system. *Hippocampus* 10:457–465.

Conitzer, V., and Sandholm, T. 2007. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning* 67:23–43.

Daw, N. D.; Niv, Y.; and Dayan, P. 2005. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neurosci.* 8:1704–1711.

Delgado, M. R. 2007. Reward-related responses in the human striatum. *Ann. New York Acad. Sci.* 1104:70–88.

Drineas, P.; Mahoney, M. W.; and Muthukrishnan, S. 2006. Sampling algorithms for l2 regression and applications. In *Proc. 17-th Annual SODA*, 1127–1136.

Franzius, M.; Sprekeler, H.; and Wiskott, L. 2007. Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology* (8). doi:10.1371/journal.pcbi.0030166.

Guestrin, C.; Koller, D.; Gearhart, C.; and Kanodia, N. 2003. Generalizing plans to new environments in relational MDPs. In *18th Int. Joint Conf. on Artif. Intell.*

Gyenes, V.; Bontovics, A.; and Lőrincz, A. 2008. Factored temporal difference learning in the New Ties environment. *Acta Cybernetica*. (under revision).

Henze, D. A.; Wittner, L.; and Buzsáki, G. 2002. Single granule cells reliably discharge targets in the hippocampal CA3 network in vivo. *Nature Neuroscience* 5:790–795.

Kearns, M., and Singh, S. 1998. Near-optimal reinforcement learning in polynomial time. In *15th Int. Conf. on Machine Learning*, 260–268. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Kloosterman, F.; van Haeften, T.; and da Silva, F. H. L. 2004. Two reentrant pathways in the hippocampal-entorhinal system. *Hippocampus* 14:1026–1039.

Koller, D., and Parr, R. 2000. Policy iteration for factored mdps. In *Proc. of the 16th Conf. on Uncertainty in Artif. Intell.*, 326–334.

Lőrincz, A., and Buzsáki, G. 2000. Two–phase computational model training long–term memories in the entorhinal–hippocampal region. *Ann. New York Acad. Sci.* 911:83–111.

Lőrincz, A., and Szabó, Z. 2007. Neurally plausible, non-combinatorial iterative independent process analysis. *Neurocomputing* 70:1569–1573.

Lőrincz, A.; Szirtes, G.; Takács, B.; and Buzsáki, G. 2001. Independent component analysis of temporal sequences forms place cells. *Neurocomputing* 38:769–774.

Lőrincz, A.; Kiszlinger, M.; and Szirtes, G. 2008. Model of the hippocampal formation explains the coexistence of grid cells and place cells. http://arxiv.org/abs/0804.3176.

Lőrincz, A.; Szatmáry, B.; and Szirtes, G. 2002. Mystery of structure and function of sensory processing areas of the neocortex: A resolution. *J. Comp. Neurosci.* 13:187–205.

Lőrincz, A. 1998. Forming independent components via temporal locking of reconstruction architectures: A functional model of the hippocampus. *Biol. Cyb.* 75:37–47.

McNaughton, B. L.; Battaglia, F. P.; Jensen, O.; Moser, E. I.; and Moser, M. 2006. Path integration and the neural basis of the cognitive map. *Nature Rev. Neuro.* 7:663–678.

Póczos, B.; Szabó, Z.; Kiszlinger, M.; and Lőrincz, A. 2007. Independent process analysis without a priori dimensional information. *Lect. Notes in Comp. Sci.* 4666:252–259.

Redish, A. D.; Battaglia, F. P.; Chawla, M. K.; Ekstrom, A. D.; Gerrard, J. L.; Lipa, P.; Rosenzweig, E. S.; Worley, P. F.; Guzowski, J. F.; McNaughton, B. L.; and Barnes, C. A. 2001. Independence of firing correlates of anatomically proximate hippocampal pyramidal cells. *Journal of Neuroscience* 21:1–6.

Schultz, W. 2002. Getting formal with dopamine and reward. *Neuron* 36:241–263.

Scoville, W. B., and Milner, B. 1957. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatry* 20:11–21.

Squire, L. R. 1992. Memory and hippocampus: a synthesis of findings with rats, monkeys and humans. *Psychol. Rev.* 99:195–231.

Sutton, R.; Szepesvári, C.; Geramifard, A.; and Bowling, M. 2008. Dyna-style planning with linear function approximation and prioritized sweeping. In *Uncertainty in Artifical Intelligence*. in press.

Szabó, Z.; Póczos, B.; and Lőrincz, A. 2007. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research* 8:1063–1095.

Szita, I., and Lőrincz, A. 2007. Learning to play using low-complexity rule-based policies: Illustrations through Ms. Pac-Man. *J. of Artif. Intell. Res.* 30:659–684.

Szita, I., and Lőrincz, A. 2008. Factored value iteration converges. http://arxiv.org/abs/0801.2069.

Szita, I.; Gyenes, V.; and Lőrincz, A. 2006. Reinforcement learning with echo state networks. *Lecture Notes in Computer Science* 4131:830–839.

Szita, I.; Takács, B.; and Lőrincz, A. 2003. ε-MDPs: Learning in varying environments. *J. of Mach. Learn. Res.* 3:145–174.

Takács, B., and Lőrincz, A. 2006. Independent component analysis forms place cells in realistic robot simulations. *Neurocomputing* 69:1249–1252.

Tesauro, G. 1995. Temporal difference learning and TD-gammon. *Communications of the ACM* 38(3):58–68.