Spike-based Reinforcement Learning in Continuous State and Action Space

Eleni Vasilaki, Robert Urbanczik, Walter Senn, Wulfram Gerstner

Computational Intelligence Seminar, June, 2009
presented by Dejan Pecevski
Outline

• Reward-based learning
• Morris water maze task
• The model
  ▪ Model architecture
  ▪ Neuron model and the learning rule
• Results
  ▪ Speed of learning
  ▪ Performance dependence on number of neurons
• Bias in the learning rule
• Summary
Reward-based Learning

- Roots in behavioral psychology
  - Instrumental conditioning: Thorndike’s cat puzzle box experiments

Thorndike’s law of effect:

*Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal...will, other things being equal, be more firmly connected with the situation...; those which are accompanied or closely followed by discomfort...will have their connections with the situation weakened...*  
(Thorndike, 1911, p. 244)

How are behavioral changes on a macroscopic level implemented by changes on the microscopic neural level?
Hebbian learning and Dopamine

- Synaptic efficacy changes according to the Hebb’s principle

\[ \Delta w_{ij} = \alpha(w_{ij}) f_1(\text{pre}_j) f_2(\text{post}_i) \]

- Local rule - basis for unsupervised and developmental learning

- Demonstrated experimentally in biological neurons

- Dopamine
  - Shown to stabilize long-term potentiation and long-term depression
  - Firing of midbrain dopaminergic neurons encodes reward prediction error

- Reward-modulated Hebbian learning

\[ \Delta w_{ij} = \alpha(w_{ij}) R f_1(\text{pre}_j) f_2(\text{post}_i) \]
Theoretical work on 
Reward-based spike-timing Hebbian plasticity

• Optimality based models
  ▪ Learning rules perform reward maximization by gradient ascent
  ▪ related to policy gradient methods in RL

• Phenomenological models
  ▪ Derive the rule based on experimental data
  ▪ Investigate the functional properties of the derived rule.

The optimality based and phenomenological models have a similar form.
Morris Water Maze Task

- Standard paradigm for behavioral learning and navigation.
- Rat (Mouse) is placed in a pool of milky (non-transparent) water.
- It has to find an invisible platform, just below the surface.
- On each trial the rat starts from a different position.
- The platform is always at a fixed position.
Model Architecture: Place cells

- 361 place cells
- Arranged in a 19x19 grid

- Poisson neurons with a firing rate
  \[ R_i^f(x, y) = r_0 \exp \left( -\frac{(x - x_i)^2 + (y - y_i)^2}{2\sigma^2} \right) \]
  
  \((x_i, y_i)\) - location where neuron i has the strongest response

- The environment is 100x100cm.

- Distance between centers of two place fields is 5 cm, width of place fields is 8 cm.
Model Architecture: Action Cells

- The place cells project to 360 action cells, arranged in a ring structure.
- Each action cell has a preferred direction \( \theta_i \).
- The action cells have Gaussian profile lateral connectivity
  \[
  w_{ij}^{lc} = w_E \exp \left( -\frac{|\theta_i - \theta_j|^2}{2\sigma^2} \right) - w_I - w_0
  \]
  \( w_{ij}^{lc} \) - weight of the lateral connection between cells i and j
- Only connections from the place cells are plastic.
Model Architecture: Generating Actions

- Each 200 ms the activity of the action cells is reset, and an action is generated.

- The angle of movement is calculated from the population vector of the action cells

\[ \theta = \arctan \left( \frac{\sum_i r_i \cos(2\pi i/N^{AC})}{\sum_i r_i \sin(2\pi i/N^{AC})} \right) \]

where \( \dot{r}_i = \frac{r_i}{\tau} + Y_i(t) \) is the low-pass filtered spiking activity of cell i.

\( N^{AC} \) is the number of action cells.
Neuron model

- Poisson Neuron model – special case of the spike response model

\[
u_i(t) = u_{rest} + \sum_{j=1}^{N} \xi_{i,j}^{f} \sum_{t_{j}^{f} \in x_j} \varepsilon(t - t_j^{f}) + \sum_{t_{i}^{f} \in y_i, l} \eta(t - t_i^{f})
\]

- Stochastic synapses: \( \xi_{i,j}^{f} \) stochastic binary variable
  - Takes value of 1 with probability \( q_{ij} \) at each presynaptic spike

- Spikes are generated by an inhomogeneous Poisson process of rate

\[
\rho_i(t) = g(u_i(t)) = \rho_0 \exp \left( \frac{u_i - u_0}{\Delta u} \right)
\]
The learning rule

- Performs gradient ascent on the average received reward (optimality based)

- Synaptic weight changes

\[
\frac{dq_{ij}(t)}{dt} = \alpha R(t)e_{ij}(t)
\]

\[0.15 \leq q_{ij} \leq 1\]

- The eligibility trace \(e_{ij}(t)\) has the following dynamics

\[
\frac{de_{ij}}{dt} = -\frac{e_{ij}}{\tau_e} + \left[Y_i(t) - \frac{\rho_i(t)}{1 + \tau_c \rho_i(t)}\right] \sum_{t_j^f \in x_{ij}} \epsilon(t - t_j^f)
\]

- \(Y_i(t) = \sum_f \delta(t - t_j^f)\) are the output spikes of neuron I

- \(\epsilon(t - t_j^f)\) - postsynaptic potential due to a presynaptic spike at \(t_j^f\)

- \(\rho_i(t)\) - instantaneous firing rate of neuron i
Mapping the model to the rat’s nervous system

- Two distinct neural systems in rats involved in different types of navigation
  1. **Taxon navigation**
     - stimulus-response chaining
     - active when the target is visible
     - includes dorsolateral striatum
  2. **Locale navigation**
     - for invisible targets
     - involves hippocampus and ventral striatum
     - relationship among events – cognitive map

- Place cells represent the “place” cells in hippocampus
  - Cells in the hippocampus exhibit high firing rates when the animal is in a specific location in the environment.

- Action cells can be located in nucleus accumbens, part of the ventral striatum.
Simulations details

• **Multiple sets of trials are performed**, where in each new set the position of the goal is changed and the synaptic weights are reinitialized.

• The target platform is at a fixed position in all trials in one trial set.

• At the beginning of a trial the rat starts from a different position near the walls.

• Positive reward is given when the rat reaches the goal (target platform).

• Negative reward is given when the rat hits a wall.

• A trial ends when the rat reaches the platform, or after 90 seconds whichever comes first.
• Escape latency – time until reaching the target.
• Results are averaged for 10 sets of trials.
• The rat learns to reach the target within 20 trials.
• Performance similar to the experimental data.
Navigation map

Arrows represent a vector sum of the preferred directions of the action cells weighted by the synaptic strengths connecting from the particular place cell.
Performance dependence on number of cells

- Average learning time does not increase with larger number of place cells.

- Overlapping place fields → many synapses are learning simultaneously

- Average learning time does not increase with larger number of action cells.

- The lateral connectivity reduces the space of possible activity profiles.

- Cells in the activity bump learn simultaneously.
Bias towards active action cells

\[
\frac{dq_{ij}(t)}{dt} = \alpha R(t)e_{ij}(t) \quad \frac{de_{ij}}{dt} = -\frac{e_{ij}}{\tau_e} + \left[ Y_i(t) - \frac{\rho_i(t)}{1 + \tau_c \rho_i(t)} \right] \sum_{t'_j \in x_{ji}} \epsilon(t - t'_j)
\]

slow learning:

- The original derived rule from reward maximization has \( \tau_c = 0 \).
- If \( \tau_c = 0 \), for the averaging to work one has to use a very small learning rate \( \alpha \).
- For larger learning rates, the fluctuations of the eligibility trace give a large noisy contribution to the weight change.
Bias towards active action cells

**solution:**
A new factor that increases the eligibility traces for active action cells in the activity bump.
The weights will change for the action neurons that generated the action before the reward.

\[ \tau_c = 0 \text{ ms} \]

\[ \tau_c = 5 \text{ ms} \]
Simulations without bias

Learning does not stabilize, due to large noisy fluctuations in the eligibility traces.
Relations to other reinforcement learning methods

- Shares properties with other policy gradient methods
  - For discrete time has the same form as Associative Reward Inaction, a REINFORCE algorithm.
  - The correlation of the pre-synaptic and posts-synaptic activity with the reward drives the learning.
  - Only associations within the time window in the eligibility trace are learned.
  - In contrast to TD methods, does not use value functions, does not bootstrap.

- Increase of the number of neurons does not increase learning time
  - provided that the width of place fields and width of lateral interactions stay the same.
Summary

- A spike-based three-factor learning rule was derived based on gradient ascent of the average reward.

- The learning rule was used in a model of learning in the neural pathways in the rat involved in locale navigation: hippocampus – ventral striatum.

- The model was tested on the Morris water maze task, a standard paradigm for behavioral learning and navigation.

- The results show that the model reproduces behavioral data of real rats in terms of escape latency vs. learning time.