Long Term Potentiation, Navigation & Dynamic Progamming

Peter Dayan
CBCL
E25-210, MIT
Cambridge, MA 02139
dayan@psyche.mit.edu

Satinder Pal Singh
Harlequin, Inc
1 Cambridge Center
Cambridge, MA 02142
singh@harlequin.com

Abstract

Blum and Abbott (1995) recently proposed an algorithm for learned navigation that is based on Hebbian changes to adaptive connections between place cells in the hippocampus. This paper suggests using a temporal difference rule (Sutton, 1988) for synaptic plasticity instead, and shows how this alters the resulting behaviour. It also recasts the problem of navigation into the reinforcement learning context of adaptive optimising control, and shows how to learn new trajectories that are not based only on averaging old ones.

Preference: Oral
Category: Theory and Analysis; Learning and Memory

Long Term Potentiation, Navigation & Dynamic Progamming

Peter Dayan
CBCL
E25-210, MIT
Cambridge, MA 02139
dayan@psyche.mit.edu

Satinder Pal Singh
Harlequin, Inc
1 Cambridge Center
Cambridge, MA 02142
singh@harlequin.com

1 Introduction

Blum & Abbott (1995) (henceforth BA; see also Abbott & Blum, 1995) considered the problem of learning about sequential behaviour in such tasks as navigating to a goal in a maze (Morris, 1981). They treated the case in which a rat has a distributed representation of its current state $\mathbf{x}$ in the activities $\{r_i(\mathbf{x})\}$ of a set of place cells whose centres are at $\{\mathbf{s}_i\}$ and whose firing profiles start by being roughly Gaussian:

$$r_i(\mathbf{x}) = f_i(\mathbf{x}) \equiv e^{-|\mathbf{x}-\mathbf{s}_i|^2/2\sigma^2} \tag{1}$$

The rat has the wherewithal to extract $\mathbf{x}$ from this group of place cells using a weighted-average scheme as:[1]

$$\hat{\mathbf{x}} = \frac{\sum_i \mathbf{s}_i r_i(\mathbf{x})}{\sum_j r_j(\mathbf{x})} \sim \mathbf{x} \tag{2}$$

Place cell i is also connected to place cell j through changeable weight ΔW_{ij}, and BA noted that changes to synaptic efficacies caused by long term potentiation require a particular temporal asymmetry – the activation of pre-synaptic fibres should generally precede that of postsynaptic fibres (Levy & Steward, 1983; Gustafsson *et al*, 1987). Based on the rat's path $\mathbf{z}(t)$ through the world, the weights are given by:

$$\Delta W_{ij} = \int H(t') f_i(\mathbf{z}(t+t')) f_j(\mathbf{z}(t)) dt \; dt' \tag{3}$$

where $H(t')$ models the timecourse for LTP. This implies that the order in which states are traversed is reflected in asymmetric coupling strengths between the cells mediating the representation. BA took these connections as exerting influence over the firing of the representing cells:

$$r'_i(\mathbf{x}) = f_i(\mathbf{x}) + \sum_j \Delta W_{ij} f_j(\mathbf{x}). \tag{4}$$

BA show that if the rat uses $r'_i(\mathbf{x})$ in equation 2 to decode its location, *without taking into account the fact that $\mathbf{s}_i$ is now incorrect,* it would believe itself to be at $\mathbf{p}(\mathbf{x})$ rather than $\mathbf{x}$, where, to first order in ΔW_{ij}

$$\mathbf{p}(\mathbf{x}) = \mathbf{x} + \frac{\sum_{ij}(\mathbf{s}_i - \mathbf{x})\Delta W_{ij} f_j(\mathbf{x})}{\sum_i f_i(\mathbf{x})} \tag{5}$$

---

* Preference: Oral
Category: Theory and Analysis; Learning and Memory

[1] BA also treat a different method for extracting x based on a maximum likelihood principle, but show that the results using it are very similar to those based on the simpler average.

Using some fairly mild approximations, BA showed further that:

$$\mathbf{p}(\mathbf{x}) \simeq \mathbf{x} + \mathcal{K} \int [\mathbf{z}(t) - \mathbf{x} + \tau \dot{\mathbf{z}}(t)]\, e^{-|\mathbf{z}(t)-\mathbf{x}|^2/4\sigma^2} dt \tag{6}$$

where $\mathcal{K}$ is a constant and τ is a characteristic time constant for LTP that is determined by $H(t)$. BA modeled the case in which the animal can use the difference $\mathbf{p}(\mathbf{x}) - \mathbf{x}$ to control its actions. The exponential factor $e^{-|\mathbf{z}(t)-\mathbf{x}|^2/4\sigma^2}$ weighs how close the trajectory of the rat came to point $\mathbf{x}$. The term depending on $\dot{\mathbf{z}}(t)$ forces the rat to move in the direction of the averaged trajectory near to $\mathbf{x}$. In maze-like tasks, BA ensured that this averaging took place only over those trajectories that resulted in the animal reaching the goal. This implies that the average trajectory will often result in the animal reaching the goal too, and, in simple tasks, reaching it more efficiently. The term depending on $\mathbf{z}(t) - \mathbf{x}$ tends to force the rat to approach places that it frequently visits. In the context of achieving a goal, this can be suboptimal – it can even pull the rat *backwards* along trajectories as well as forwards.

2 Temporal Difference Learning

Consider replacing the update rule in equation 3 by:

$$\Delta W_{ij} = \int dt\; dt'\, H(t') \left(f_i(\mathbf{z}(t+t')) - f_i(\mathbf{z}(t))\right) f_j(\mathbf{z}(t)). \tag{7}$$

using the *difference* in postsynaptic activity rather than its absolute level. Montague & Sejnowski (1994) discuss the neurobiological and computational bases of a version of this rule – its use here is inspired by the temporal difference rules (Sutton, 1988) that we describe in the next section.

Making learning sensitive to the averaged temporal difference in postsynaptic activity rather than its absolute level has two beneficial effects for BA. First, in the continuum limit, it preserves the sum total activity $\sum_i r_i(\mathbf{x})$. Since this expression is the denominator in equation 2, it is required to turn the activities $r_i(\mathbf{x})$ into estimates of location, and so it is desirable that it remain fixed.

We show this in the continuum case studied by BA in which sums over i are converted into integrals over the two-dimensional $d\mathbf{s}_1$, sums over j into integrals over the two-dimensional $d\mathbf{s}_2$, both representing the centres of the place fields, ΔW_{ij} is converted into $\Delta W(\mathbf{s}_1 - \mathbf{s}_2)$, and

$$f_i(\mathbf{x}) \sim e^{-|\mathbf{x}-\mathbf{s}_1|^2/2\sigma^2}.$$

Then $\sum_i r'_i(\mathbf{x}) - \sum_i r_i(\mathbf{x})$

$$= \sum_{ij} \Delta W_{ij} f_j(\mathbf{x}) \tag{8}$$

$$\simeq \int \Delta W(\mathbf{s}_1 - \mathbf{s}_2) e^{-|\mathbf{x}-\mathbf{s}_2|^2/2\sigma^2} d\mathbf{s}_1 d\mathbf{s}_2 \tag{9}$$

$$= \int H(t') \left(e^{-|\mathbf{z}(t+t')-\mathbf{s}_1|^2/2\sigma^2} - e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}\right) e^{-|\mathbf{z}(t)-\mathbf{s}_2|^2/2\sigma^2} e^{-|\mathbf{x}-\mathbf{s}_2|^2/2\sigma^2} d\mathbf{s}_1 d\mathbf{s}_2 dt dt' \tag{10}$$

$$
\begin{aligned}
&\simeq \int H(t')t'\dot{\mathbf{z}}(t)\cdot(\mathbf{s}_1-\mathbf{z}(t))e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}e^{-|\mathbf{z}(t)-\mathbf{s}_2|^2/2\sigma^2}e^{-|\mathbf{x}-\mathbf{s}_2|^2/2\sigma^2}d\mathbf{s}_1d\mathbf{s}_2dtdt' &(11)\\
&= 0 &(12)
\end{aligned}
$$

where equation 11 follows on making the approximations that $H(t')$ is small for t' large and the animal does not move too fast and that therefore:

$$
\begin{aligned}
e^{-|\mathbf{z}(t+t')-\mathbf{s}_1|^2/2\sigma^2}-e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2} &\simeq (\mathbf{z}(t+t')-\mathbf{z}(t))\cdot\nabla_{\mathbf{z}(t)}e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}\\
&\simeq t'\dot{\mathbf{z}}(t)\cdot(\mathbf{s}_1-\mathbf{z}(t))e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}
\end{aligned}
$$

and equation 12 follows on performing the integral

$$\int(\mathbf{s}_1-\mathbf{z}(t))e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}d\mathbf{s}_1=0.$$

Therefore $\sum_i r_i(\mathbf{x})$ remains constant.

The second and more interesting desirable outcome from using equation 7 for setting the weights is that the term in equation 6 that depends on $\mathbf{z}(t)-\mathbf{x}$ is eliminated. The same arguments that led to equation 12, imply that
$\mathbf{p}(\mathbf{x})-\mathbf{x}$:

$$
\begin{aligned}
&\propto \int(\mathbf{s}_1-\mathbf{x})\Delta W(\mathbf{s}_1-\mathbf{s}_2)e^{-|\mathbf{x}-\mathbf{s}_2|^2/2\sigma^2}d\mathbf{s}_1d\mathbf{s}_2 &(13)\\
&\simeq \int(\mathbf{s}_1-\mathbf{x})H(t')t'\dot{\mathbf{z}}(t)\cdot(\mathbf{s}_1-\mathbf{z}(t))e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}e^{-|\mathbf{z}(t)-\mathbf{s}_2|^2/2\sigma^2}e^{-|\mathbf{x}-\mathbf{s}_2|^2/2\sigma^2}d\mathbf{s}_1d\mathbf{s}_2dtdt' &(14)\\
&\propto \int(\mathbf{s}_1-\mathbf{x})H(t')t'\dot{\mathbf{z}}(t)\cdot(\mathbf{s}_1-\mathbf{z}(t))e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}e^{-|\mathbf{z}(t)-\mathbf{x}|^2/4\sigma^2}d\mathbf{s}_1dtdt' &(15)\\
&\propto \int(\mathbf{s}_1-\mathbf{x})\dot{\mathbf{z}}(t)\cdot(\mathbf{s}_1-\mathbf{z}(t))e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}e^{-|\mathbf{z}(t)-\mathbf{x}|^2/4\sigma^2}d\mathbf{s}_1dt &(16)\\
&\propto \int\dot{\mathbf{z}}(t)e^{-|\mathbf{z}(t)-\mathbf{x}|^2/4\sigma^2}dt &(17)
\end{aligned}
$$

where equation 15 follows on integrating out $\mathbf{s}_2$, equation 16 on integrating $H(t')t'$ over t', and equation 17 since:

$$\int(\mathbf{s}_1-\mathbf{x})_a(\mathbf{s}_1-\mathbf{z}(t))_b e^{-|\mathbf{z}(t)-\mathbf{s}_1|^2/2\sigma^2}d\mathbf{s}_1\propto\delta_{ab}$$

where δ_{ab} is the Kronecker-delta.

Comparing equations 6 and 17, we can see that this way of defining ΔW implies that $\mathbf{p}(\mathbf{x})-\mathbf{x}$ points exactly in the direction of the averaged neighbouring velocity, eliminating the unwanted component.

3 Optimal Control

Averaging the velocity over the trajectories that result in the animal attaining the goal should be enough to guarantee that it can return there using this method of navigation, at least provided that the space is not too tortuous and the sampled trajectories cover enough of it. Indeed, BA show empirically the success of their learning rule in simple environments. However, one might also be interested in having the animal learn good paths to the goal, not just any path to the goal. The substrate for continual improvement certainly looks as if it is available. For instance, the centres $\mathbf{s}_i$ of the place cells could adapt to absorb the changes coming from ΔW_{ij}, and then new values of ΔW_{ij} could be learnt, driving the animal in a new direction, and then leading to further adaptation of $\mathbf{s}_i$, and so forth. Unfortunately, this process would not work in general. Equation 17 suggests that it is only ever possible to average over existing trajectories. For instance, consider the case in which there are multiple ways to get to the goal. If the initial actions of the animal never encompassed one of these ways, then no amount of averaging would be guaranteed to reveal it. Furthermore, if different routes to the goal cost different amounts in a metric other than time (*eg* one demands more effort), then it would be convenient to have some way of favouring cheap routes over expensive ones.

An appropriate framework in which to analyse optimising behaviour is the algorithm called policy improvement, which is Howard's (1960) method for finding optimal policies in MDPs. This formalises the problem in a slightly different way. Consider the task of controlling a deterministic system to maximise

$$V^*(\mathbf{x}_0) = \max_{\mathbf{u}(t,\mathbf{z}(t))} \int_0^\infty q(\mathbf{x}(t), \mathbf{u}(t, \mathbf{x}(t)))dt \tag{18}$$

where $\mathbf{z}(t) \in \Re^n$ is the state at time t, $\mathbf{u}(t, \mathbf{z}(t)) \in \Re^m$ is the control, $\mathbf{z}(0) = \mathbf{x}_0$, $\dot{\mathbf{z}}(t) = \mathbf{h}(\mathbf{z}(t), \mathbf{u}(t, \mathbf{z}(t)))$, and $q(\mathbf{x}, \mathbf{u})$ is the scalar return or cost for performing action $\mathbf{u}$ at state $\mathbf{x}$. A simple model of a maze task would have $0 > q(\mathbf{z}, \mathbf{u}) = -\kappa$ and $\dot{\mathbf{z}}(t) = \mathbf{u}(\mathbf{z}(t))$ everywhere away from the goal, where we may assume that $|\mathbf{u}|^2$ is fixed so that the animal maintains a constant speed. $V^*(\mathbf{x}_0)$ is called the optimal value function, and in this case, assigns to states $\mathbf{x}_0$ a value proportional to minus the minimal length of time it could take the animal to get to the goal from $\mathbf{x}_0$. Minimising the time is the aim – however, different values for q could be used, for instance to model extra penalties for being in particular parts of the maze.

In the complete version of this paper, we show that we can take a *policy* $\mathbf{w}(\mathbf{x})$, which assigns controls to states, and learn a the value function for policy $\mathbf{w}$, $V^{\mathbf{w}}(\mathbf{x})$, which estimates the worth of state $\mathbf{x}$ under policy $\mathbf{w}$. $V^{\mathbf{w}}(\mathbf{x})$ is represented as:

$$V^{\mathbf{w}}(\mathbf{x}) = \sum_i c_i^{\mathbf{w}} r_i(\mathbf{x}), \tag{19}$$

with parameters $c_i^{\mathbf{w}}$ and therefore requires a set of adaptive connections between the place cells and a neuron (or more likely, a collection of neurons) that learns $V^{\mathbf{w}}(\mathbf{x})$. It is a form of radial basis function representation for $V^{\mathbf{w}}(\mathbf{x})$ (Broomhead & Lowe, 1988; Poggio & Girosi, 1990).

With $r_i(\mathbf{x}) = f_i(\mathbf{x})$, $c_i^{\mathbf{w}}$ are updated according to:

$$\Delta c_j^{\mathbf{w}} \propto \left[-\kappa + V^{\mathbf{w}}(\mathbf{x} + \epsilon\mathbf{w}(\mathbf{x})) - V^{\mathbf{w}}(\mathbf{x})\right] f_j(\mathbf{x}) \tag{20}$$

where the $-\kappa$ stands in for the negative reinforcement coming from swimming being aversive to the rat. This is again a form of temporal difference rule. In this case of constant reinforcement, a method akin to BA's can be used to calculate the new direction for the rat to move that comes from one step of Howard's policy improvement algorithm. We show that if the firing of each place cell is increased by the magnitude of its contribution to $V^{\mathbf{w}}(\mathbf{x})$, namely $c_i^{\mathbf{w}} f_i(\mathbf{x})$, then the direction $\mathbf{p}(\mathbf{x}) - \mathbf{x}$ calculated in the same way as BA will point in the direction of the improved policy.

Therefore, using a separate set of connections to learn the value function, and changing the activities of the place cells according to their contributions to $V^{\mathbf{w}}(\mathbf{x})$ rather than their links with each other, BA's system can be made to perform policy improvement and therefore not only average trajectories over ones that have been successful, but actually find successively better trajectories.

References

[1] Abbott, LF & Blum, KI (1995). Functional significance of long-term potentiation for sequence learning and prediction. *Cerebral Cortex,* in press.

[2] Bellman, RE (1957). *Dynamic Programming.* Princeton, NJ: Princeton University Press.

[3] Blum, KI & Abbott, LF (1995). A model of spatial map formation in the the hippocampus of the rat. *Neural Computation,* in press.

[4] Broomhead, DS & Lowe, D (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems,* **2**, 321-55.

[5] Dayan, P & Singh, SP (1996). Improving policies without measuring merits. In DS Tourtetzky, MI Mozer & TK Leen, editors, *Advances in Neural Information Processing, 8.* Cambridge, MA: MIT Press.

[6] Gustafsson, B, Wigstrom, H, Abraham, WC & Huang, YY (1987). Long-term potentiation in the hippocampus using depolarizing current pulses as the conditioning stimulus to single volley synaptic potentials. *Journal of Neuroscience,* **7**, 774-780.

[7] Howard, RA (1960). *Dynamic Programming and Markov Processes.* New York, NY: Technology Press & Wiley.

[8] Levy, WB & Steward, O (1983). Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience,* **8**, 791-797.

[9] Montague, PR & Sejnowski, TJ (1994). The predictive brain: Temporal coincidence and temporal order in synaptic learning mechanisms. *Learning and Memory,* **1**, 1-33.

[10] Morris, RGM (1981). Spatial localisation does not require the presence of local cues. *Learning and Motivation,* 12, 239-260.

[11] Poggio, T & Girosi, F (1990). A theory of networks for learning. *Science,* **247**, 978-982.

[12] Sutton, RS (1988). Learning to predict by the methods of temporal difference. *Machine Learning,* **3**, pp 9-44.