Reinforcement Learning: Predicting Rewards

Image Source: Wikimedia Commons

Reinforcement Learning

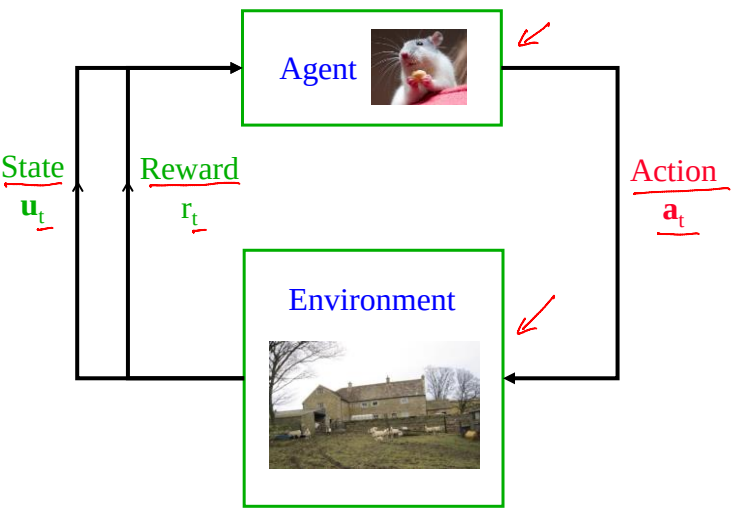


Image Source: Wikimedia Commons

Early Results: Pavlov and his Dog

- Classical (Pavlovian) conditioning experiments
- Training: Bell →Food
- After: Bell → Salivate
- Conditioned stimulus (bell) predicts future reward (food)

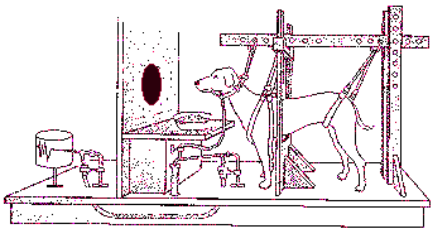


Image: Wikimedia Commons; Animation: Tom Creed, SJU

Predicting Delayed Rewards

- How do we predict rewards delivered *some time after* a stimulus is presented?
- Given: Many trials, each of length T time steps
- Time within a trial: $0 \le t \le T$ with stimulus $u(t)$ and reward $r(t)$ at each time step t (Note: $r(t)$ can be zero for some t)
- We would like a neuron whose output $v(t)$ predicts the *expected total future reward* starting from time t

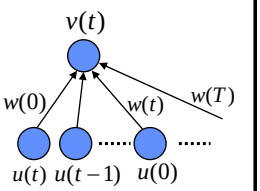
$$v(t) \approx \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) \right\rangle_{trials}$$

Learning to Predict Future Rewards

- Use a set of synaptic weights $w(t)$ and *predict based on all past stimuli* $u(t)$:

$$v(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau)$$

(Linear filter!)



- Learn weights $w(\tau)$ that minimize error:

$$\left(\sum_{\tau=0}^{T-t} r(t+\tau) - v(t)\right)^2$$

(Can we minimize this using gradient descent and delta rule?)

Yes, BUT future rewards are not yet available!

Temporal Difference (TD) Learning

- **Key Idea**: Rewrite error function to get rid of future terms:

$$\left(\sum_{\tau=0}^{T-t} r(t+\tau) - v(t)\right)^2 = \left(r(t) + \sum_{\tau=0}^{T-t-1} r(t+1+\tau) - v(t)\right)^2$$

$$\approx \left(r(t) + v(t+1) - v(t)\right)^2$$

t时刻得到reward，t+1时刻以后的reward，t时刻以后的reward

Minimize this using gradient descent!

- **Temporal Difference (TD) Learning**:

$$\Delta w(\tau) = \varepsilon\,[\,r(t) + v(t+1) - v(t)\,]\,u(t-\tau)$$

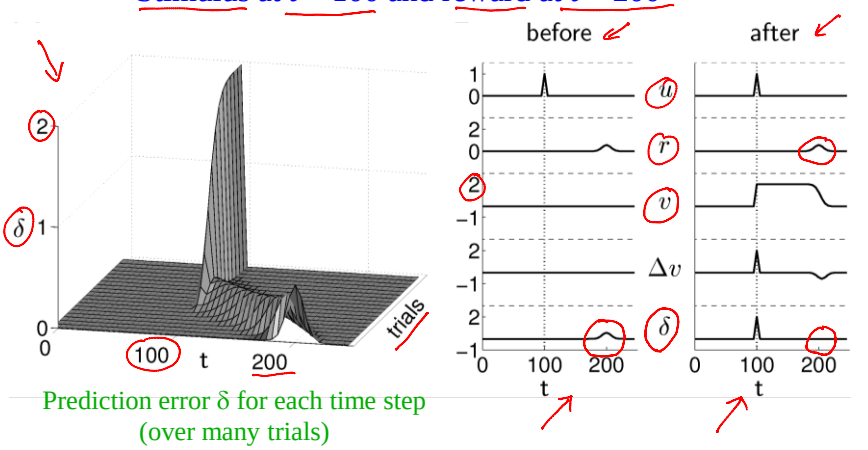
δ

Expected future reward

Prediction

Predicting Future Rewards: TD Learning

Stimulus at t = 100 and reward at t = 200



Prediction error δ for each time step
(over many trials)

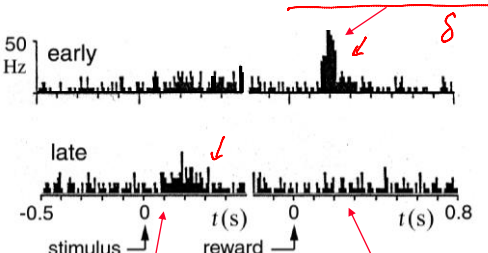
Image Source: Dayan & Abbott textbook

Possible Reward Prediction Error Signal in the Primate Brain

Dopaminergic cells in Ventral Tegmental Area (VTA)

Reward Prediction error? $[r(t)+v(t+1)-v(t)]$

就是现在的奖励



Before Training

After Training

对于这个例子

$[0+v(t)-v(t-1)]$

No error

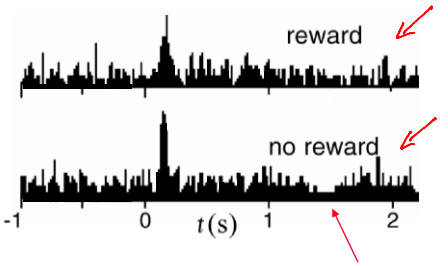
$v(t) \approx r(t)+v(t+1)$

$[r(t)+v(t+1)-v(t)] \approx 0$

Image Source: Dayan & Abbott textbook

More Evidence for Prediction Error Signals

Dopaminergic cells in VTA after Training



Negative error

$$r(t) = 0, v(t+1) = 0$$

$$[r(t) + v(t+1) - v(t)] = -v(t)$$

Reward predicted but not delivered

Image Source: Dayan & Abbott textbook

Next: Reinforcement Learning: Acting to Maximize Rewards