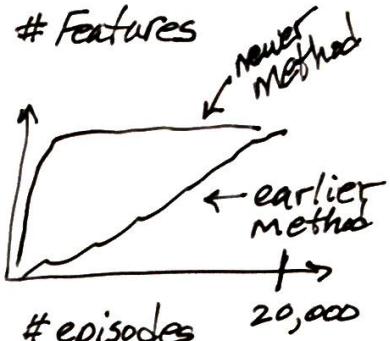
687 2017-12-04

Some discussion of homework assignment:

$$\phi(s,a) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \phi(s) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow a^{th} \text{ entry}$$

$|\theta|$ = # Actions × # Features

Typical learning plot



Natural Policy Gradient

$\tilde{\nabla} J(\theta) = w$

You pick G to adjust notion of distance: $\tilde{\nabla} J(\theta) = G(\theta)^{-1} \nabla J(\theta)$

G can be I → Euclidean distance

G can be $F(\theta) = \sum_s d^{\pi}(s) \sum_a \pi(s,a) \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta} \cdot \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}^T$

↑ Fisher information matrix

We move in policy space, irrespective of original space

NAC-TD (Natural Actor-Critic) (goes by other names)

$$\begin{aligned}
&\delta \leftarrow R_t + \gamma v^T \phi(s_{t-1}) - v^T \phi(s_t) \\
&e_v \leftarrow \gamma \lambda e_v + \phi(s_t) \\
&v \leftarrow v + \alpha_v \delta e_v \\
&e_w \leftarrow \gamma \lambda e_w + \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta} \\
&w \leftarrow w + \alpha_w \delta e_w \\
&\theta \leftarrow \theta + \alpha_\theta w
\end{aligned}$$

(TD(λ))

Can work amazingly well, BUT

- Need ϕ representation
- Need π "
- Need $\alpha_v, \alpha_w, \alpha_\theta, \lambda, \gamma$

} a LOT of hyperparameters

REINFORCE: $q^{\pi}(s,a) \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}$

high variance

Add a baseline / control: use $q^{\pi}(s,a) - v(s)$ ← baseline

Actor-critic: δ_t Then add function approximation, and

Generalized distance metric

Psychology + Neuroscience

Psych + RL

- Operant (or Instrumental) Conditioning = Learning process through which the strength of a behavior is modified by reward or punishment.

  ~ the control aspect of RL

- Classical Conditioning = Process by a biologically potent stimulus (e.g., food) is paired with a previously neutral stimulus (e.g., a bell).

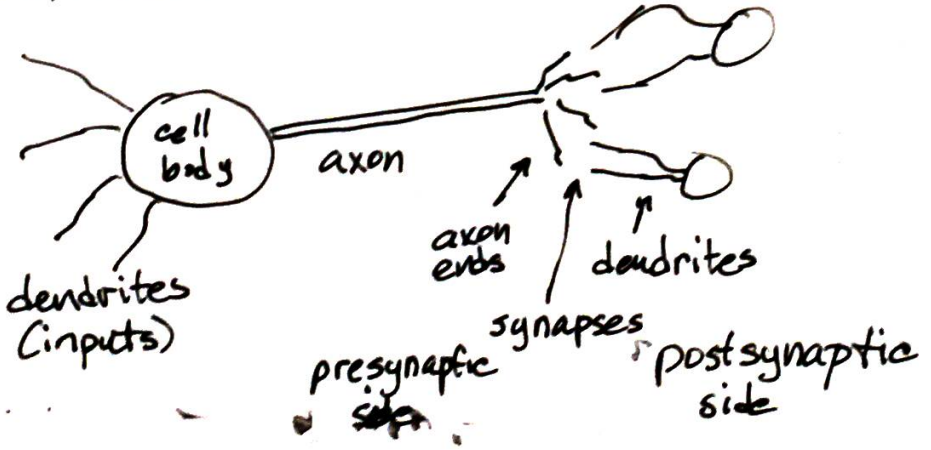
  ~ the prediction aspect of RL

Neuroscience + RL

↓

Dopamine - a neurotransmitter

Reinf. Lrning + Decision Making (RLDM) conference



Different neurons produce varying amounts of different neurotransmitters. Dopaminergic produce dopamine.

2 clusters of these in mammals

SN pc - substantia nigra pars compacta

striatum → coordinates motor + action planning, decision making, motivation

VTA - Ventral tegumental area

Many areas → prefrontal cortex: planning, personality, decision making

Olds + Milner (1954) connect dopamine w/ rewards

Barto - what if it is a TD error?

The Reward Prediction error hypothesis of dopamine neuron activity. Schultz 1997 (Science).

TED talk - optogenetics: make selected neurons fire when exposed to a flash of light.

Behavioral control loop. Actor-Critic Model

Find the Critic cells. Just 12 cells in a fruit fly!

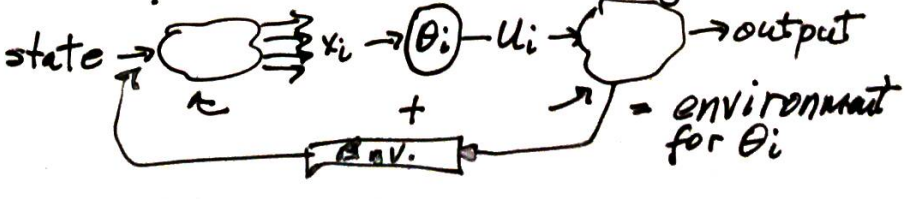
How could brains implement an RL algorithm?

A problem: Brains have no back propagation.

$$\theta \leftarrow \theta + \alpha \delta \frac{\partial \ln \pi(s,a,\theta)}{\partial \theta}$$

← how do changes to θ change things? Fundamentally backprop!

View each node as an agent:



$$\frac{\partial J(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_1} \\ \frac{\partial J(\theta)}{\partial \theta_2} \\ \vdots \end{bmatrix}$$

$\frac{\partial J(\theta)}{\partial \theta_2}$ ← objective of θ_2

network as a whole can compute the policy gradient.

$$\theta_i \leftarrow \theta_i + \alpha \delta \frac{\partial \ln \pi_i(X_i; U_i; \theta_i)}{\partial \theta_i}$$

matches the structure where the TD error is broadcast to many neurons.

See paper on:
Conjugate MDP in ICML 2011
NIPS 2011

(Note: Back prop is not easy in stochastic nets)