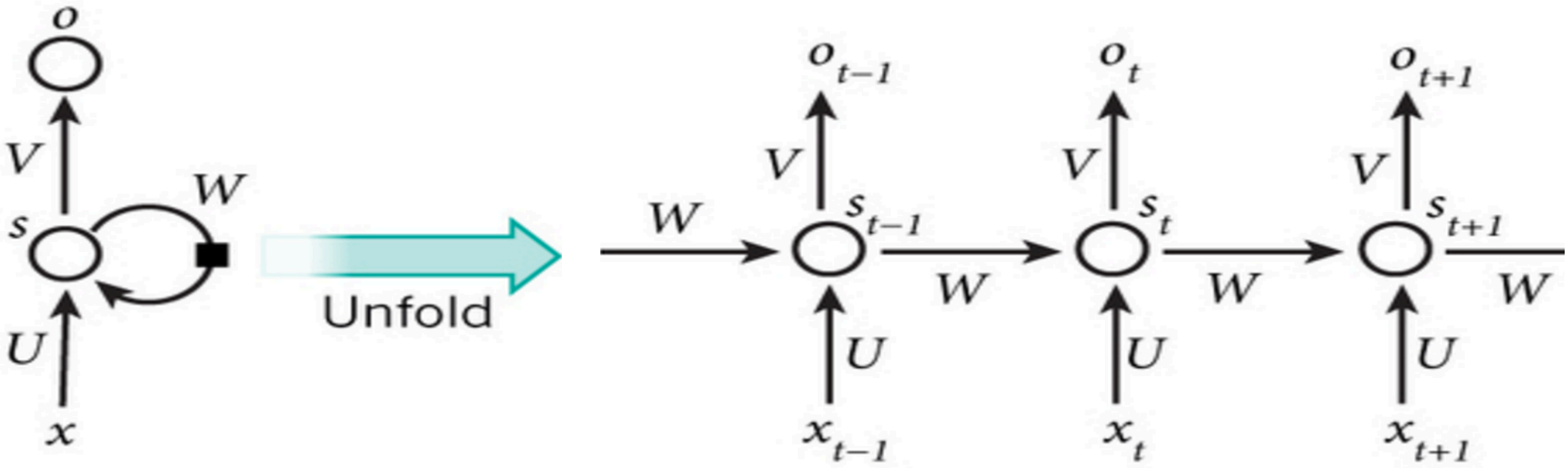
RL Recitation

Lyle Ungar

Dynamical systems: HMMs, RNNs, RL

- ???



RL Types

- **Model based**
  - Explicitly learn $p(s_{t+1}|s_t, a_t)$, $r(s_t, a_t)$
  - Markov Decision Process (MDP) or POMDP
- **Model free**
  - Learn expected value of each state, $V(s_t)$, *given a policy*
  - Learn expected value of each state and action, $Q(s_t, a_t)$
  - Learn an optimal policy, while learning V or Q
    - Can learn on- and off-policy

State can be discrete or real, *V* and *Q* can be neural nets

Notation summary

- s_t **state**
- $a_t = \pi(s_t)$ **policy π and action a_t**
- $V_\pi(s_t)$ **estimated value of s_t**
- $Q_\pi(s_t, a_t)$ **estimated value taking action a_t in s_t**
- $r(s_t, a_t, s_{t+1})$ **reward** (usually simply $r(s_t)=R_t$)
- G_t **expected discounted reward ('return')**
- γ **discount factor**
- $p(s_{t+1}|s_t, a_t)$ **model**

What is the relationship between $V(s_t)$ and G_t?

TD(0)

One could learn V(s) **by updating it at the end of each game based on who won.**

Why update *V(s)* as soon as one makes a move and sees the opponent's response?

Model based vs. Model free

- One can learn a model of the world $p(s_{t+1}|s_t,a_t)$ and use that to find an optimal policy
- Or one can learn the value $V(s_t)$ of each state - or of the action in each state $Q(s_t,a_t)$ - without a world model
- When is each better?

A problem?

- $V(s)$ **depends on** π
- **But** π **is a function of** $V(s)$

A 0.812	B 0.868	C 0.918	Food 1.00
D 0.762		E 0.660	Shock -1.00
J 0.705	G 0.655	H 0.611	I 0.388

So how can we learn $V(s)$ **and** π^***?**

Bellman Equation

Bellman's Equation: Holds for all policies $\pi(a|s)$

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right], \forall s \in \mathcal{S}$$

$$q_\pi(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_\pi(s')\right], \forall s \in \mathcal{S}, \forall a \in \mathcal{A}(s)$$

Bellman Equation (Optimality)

Bellman's Optimality Equation: Holds for optimal policies $\pi^*(s)$

$$v_*(s) = \max_{a \in \mathcal{A}(s)} \sum_{s',r} p(s', r|s, a) \left[r + \gamma v_*(s') \right], \forall s \in \mathcal{S}$$

$$q_*(s, a) = \sum_{s',r} p(s', r|s, a) \left[r + \gamma \max_{a'} q_*(s') \right], \forall s \in \mathcal{S}, \forall a \in \mathcal{A}(s)$$

Can used model-based or model-free

Bellman Equation (Q version)

$$Q(s,a) = r(s,a) + \gamma \max_{a} Q(s',a)$$

Q-Learning

$$Q(s,a) := r(s,a) + \gamma \max_{a} Q(s',a)$$

We still need to figure out how to adjust the policy - on policy or off policy

Look at each method for gridworld

- **Q values in each state**
  - State = A, B, C, ..
  - Action = l, r, u, d (left, right, up, down)

e.g.
Q(A,r) = .8

A **Q(A,r)=.8** **Q(A,d)=.7**	**B** **Q(B,l)= .75** **Q(B,r)= .9**	**C** **Q(C,l)= .8** **r=.95; d=0.7**	**Food 1**
D Q(D,u)= .8 Q(D,d)= .7	XXXXXXXX XXXXXXXX XXXXXXXX	E Q(E,u)= .8 Q(E,d)= .65	Shock -1
J Q(J,u)= .7 Q(J,r)= .65	G Q(G,l)= .7 Q(G,r)= .6	H Q(H,l)= .65 r=.6; u=0.75	I Q(I,l) = .65 Q(I,u) = -.5

Dynamic Programing, γ=1

- **Start in B, with π^*, assume**
  - p(C|B,r)=0.9 P(A|B,r)=0.1 p(C|B,l)=0.1 P(A|B,l)=0.9
- **What is the new estimate of *Q(B,r)?***

A **Q(A,r)=.8** **Q(A,d)=.7**	**B** **Q(B,l)= .75** **Q(B,r)= .9**	**C** **Q(C,l)= .8** **r=.95; d=0.7**	**Food 1**
D Q(D,u)= .8 Q(D,d)= .7	XXXXXXXX XXXXXXXX XXXXXXXX	E Q(E,u)= .8 Q(E,d)= .65	Shock -1
J Q(J,u)= .7 Q(J,r)= .65	G Q(G,l)= .7 Q(G,r)= .6	H Q(H,l)= .65 r=.6; u=0.75	I Q(I,l) = .65 Q(I,u) = -.5

Bellman's Equation

$$q_*(s,a) = \sum_{s',r} p(s',r|s,a)\left[r + \gamma \max_{a'} q_*(s')\right], \forall s \in \mathcal{S}, \forall a \in \mathcal{A}(s)$$

$$\begin{aligned} Q(B,r) &= p(C|B,r)[0+1*Q(C,r)] + p(A|B,r)[0+1*Q(A,r)] \\ &= 0.9 \quad [\quad 0.95 \quad] + \quad 0.1 \quad [\quad 0.8 \quad] \\ &= 0.935 \end{aligned}$$

TD(0) - Q-learning, γ=1, α=0.6

- Start in *s=B*, pick best action *a=r*
- Observe new state *s=C*
- What is the new estimate of *Q(B,r)?*

A Q(A,r)=.8 Q(A,d)=.7	B Q(B,l)= .75 Q(B,r)= .9	C Q(C,l)= .8 r=.95; d=0.7	Food 1
D Q(D,u)= .8 Q(D,d)= .7	XXXXXXXXX XXXXXXX XXXXXXXX	E Q(E,u)= .8 Q(E,d)= .65	Shock -1
J Q(J,u)= .7 Q(J,r)= .65	G Q(G,l)= .7 Q(G,r)= .6	H Q(H,l)= .65 r=.6; u=0.75	I Q(I,l) = .65 Q(I,u) = -.5

TD(0)

- $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left(r_t + \gamma Q\left(s_{t+1}, \pi(s_{t+1})\right) - Q(s_t, a_t)\right)$
- $Q(B, r) \leftarrow Q(B, r) + \alpha\left(0 + 1Q(C, r) - Q(B, r)\right)$
- $Q(B, r) \leftarrow 0.75 + 0.6 * (0 + 0.95 - 0.75)$
- $Q(B, r) = 0.87$