Real-time reference notes / Lecture 01

1. RL basics: States, actions, rewards, returns
2. Intuition behind value-based methods, tabular version
3. Filling up the Q-table using ϵ-greedy algorithms
4. Scalability of tabular RL
   - (a) Size of table
   - (b) Continuous states & actions
   - (c) Visiting each (s,a) pair
5. Approximation of $Q(s,a)$ using neural networks
   - (a) $s, a \rightarrow Q(s,a)$
   - (b) $s \rightarrow Q(s,a_1), Q(s,a_2), Q(s,a_3)$
6. Terminology
   - (a) Model-free & Model-based
   - (b) online & offline
   - (c) On-policy & off-policy
7. Bellman equation
   - ϵ-greedy using Bellman equation
8. MKS paper (ATARI)
9. Issues with basic DQN
   - Bootstrapping
   - stabilisation
   - Last layer discrimination
   - sampling memory

$$\dot{x} = Ax + Bu$$

$$J = \int_0^\infty (x^T Q x + u^T R u)\, dt$$

State, action

$$x_{k+1} = A^* x_k + B^* u_k$$

transition

$$J = \sum_0^\infty [x_k^T Q x_k + u_k^T R u_k]$$

cost (t_k)

$$G_t = \sum_{\tau=t}^{\infty} r_t \gamma^{\tau - t}$$

s_t , a_t

$|r_t| \le R_0$

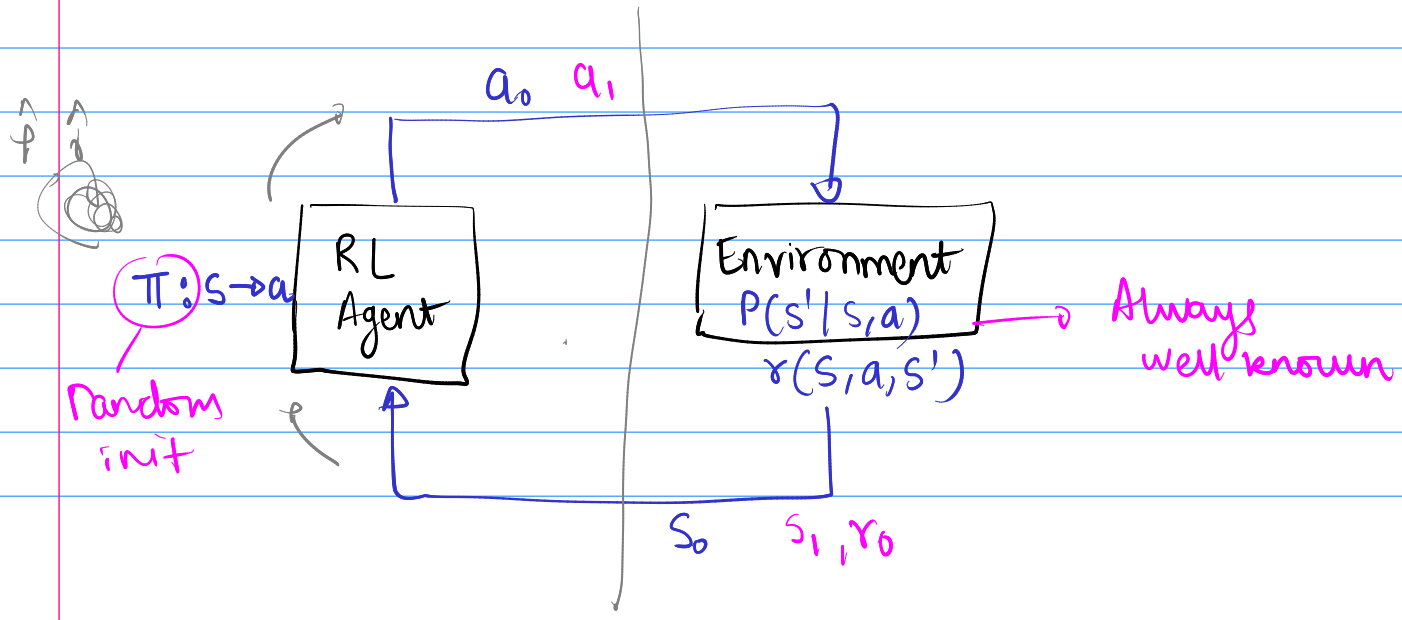
$$\max |G_t| = \frac{R_0}{1-\gamma}$$

$P(S_{t+1} \mid S_t, a_t)$

transition prob.

Markov decision process (MDP) → $S_{t-1}, a_{t-1}, S_{t-2}, \ldots, S_0, a_0$ independent

(S, a, r, P, γ) → defines MDP



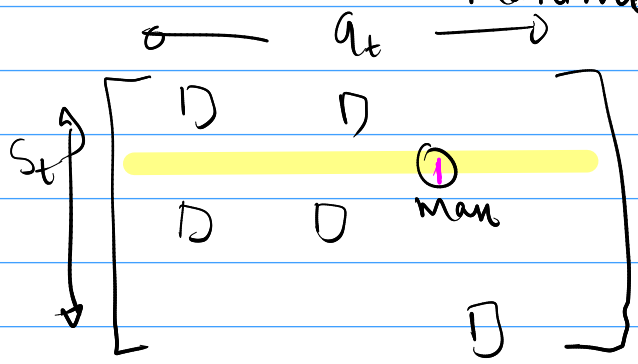
$$\max_{a_t} \boxed{G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots} = r_t + \gamma G_{t+1}$$

$\pi : s_t \to a_t$

(r_t)

$\hat{G}_t(a_t = 0)$? $\quad$ $\hat{G}_t(a_t = 1)$?

maximum



Tabular RL

ϵ-greedy exploration

- w.p. ϵ — choose randomly — exploration
- w.p. $(1-\epsilon)$ — choose argmax — exploitation

$$G_t = r_t + \gamma G_{t+1}$$

$\hat{G}$

$$Q(s_t, a_t) = r_t + \gamma Q(s_{t+1}, a_{t+1})$$

rewrite

$$\boxed{Q(s_t, a_t) = r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})}$$

Bellman equation

reward not for this regression problem

$$L = \left[Q(s_t, a_t) - r_t - \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \right]^2 \to \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \text{OC} \quad 0 \to Q$$

s_t 0 0 0 → $Q(s_t, a_t)$

a_t 0 0 0

0 0 0 → 0 → $Q_\theta(s_t, a_1) - [r_t + \gamma Q_\theta(s_{t+1})]$

0 → $Q_\theta(s_t, a_2) - Q_\theta(s_t, a_2)$