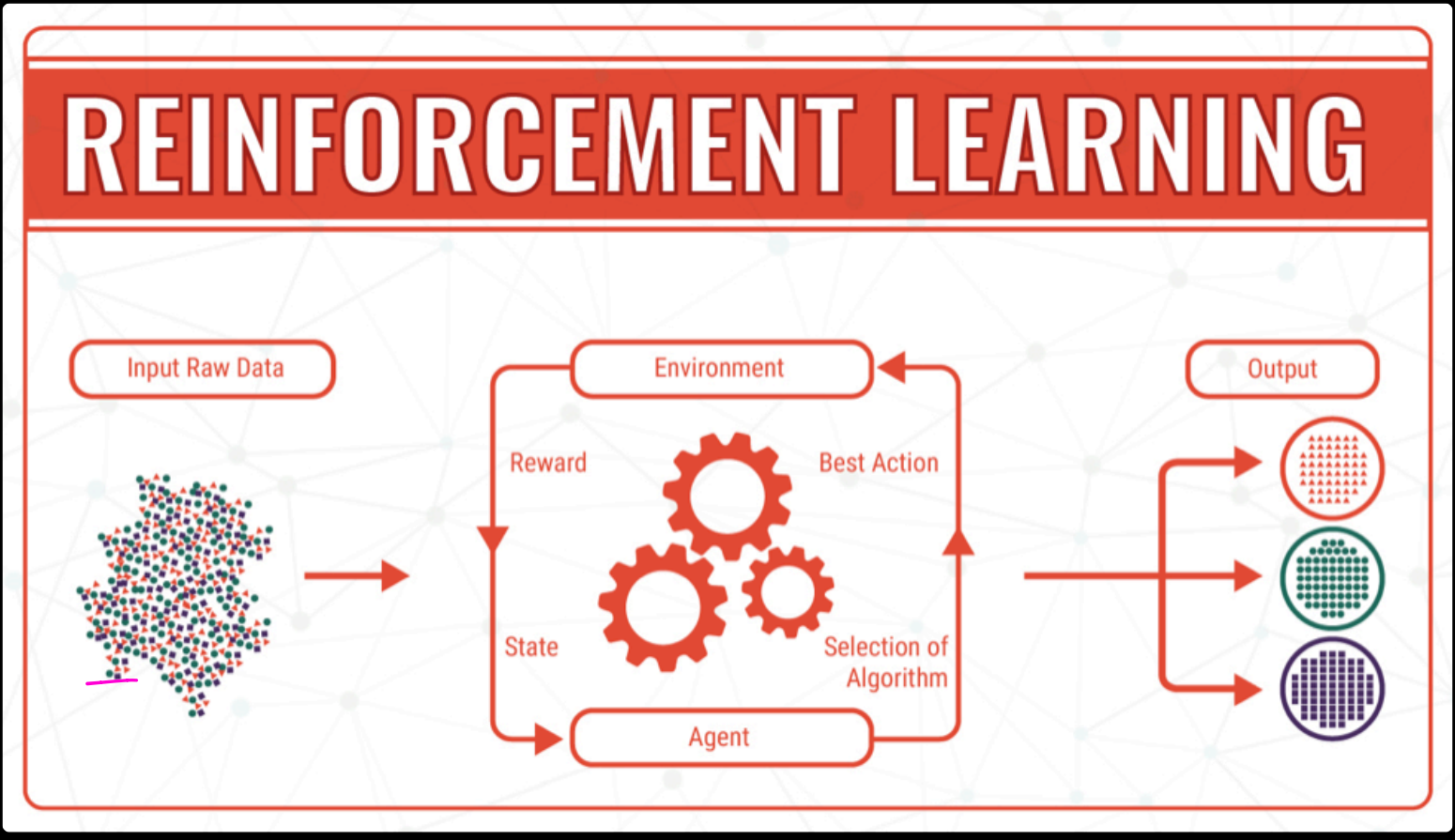
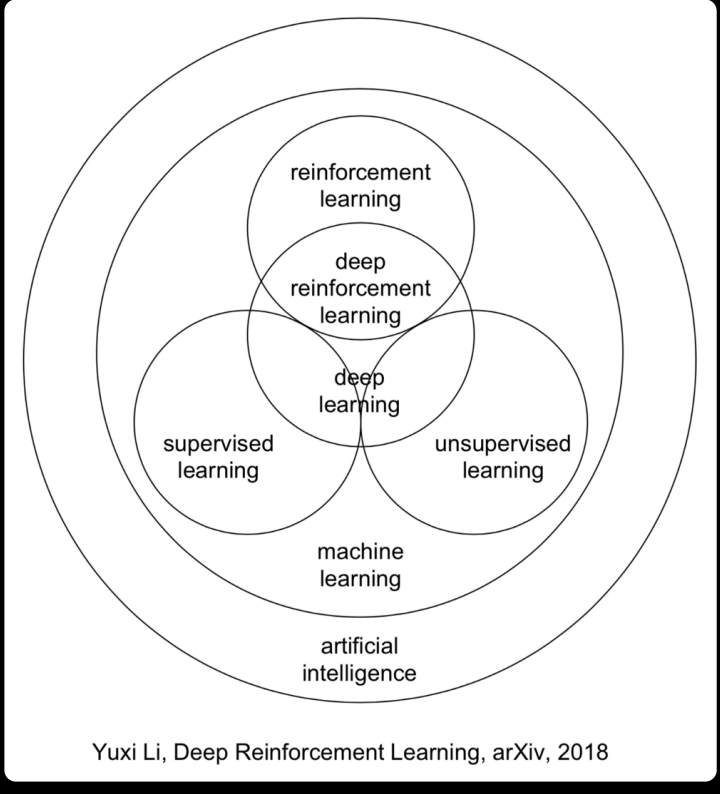
This lecture is completely adapted from Stanford lecture series: CS-231n by Serena Yeung





Yuxi Li, Deep Reinforcement Learning, arXiv, 2018

(A) → Introduction to learning paradigms

There are (3) major paradigms of m/c learning

(I) Supervised Learning

- (i) Data: (x, y)
  - x → data
  - y → label
- (ii) Goal: Learn a fn that can map $(x \rightarrow y)$

Examples: Classification (Cat Vs Dog), Object detection, Semantic segmentation, Regression $[f(x) : y]$, image Captions.

(Anything in which target is known/ given atleast for training)

(II) Unsupervised Learning

- (i) Data: (x) w/o labels
- (ii) Goal: learn some underlying hidden relationship between the data that can be used to explore the structure of data

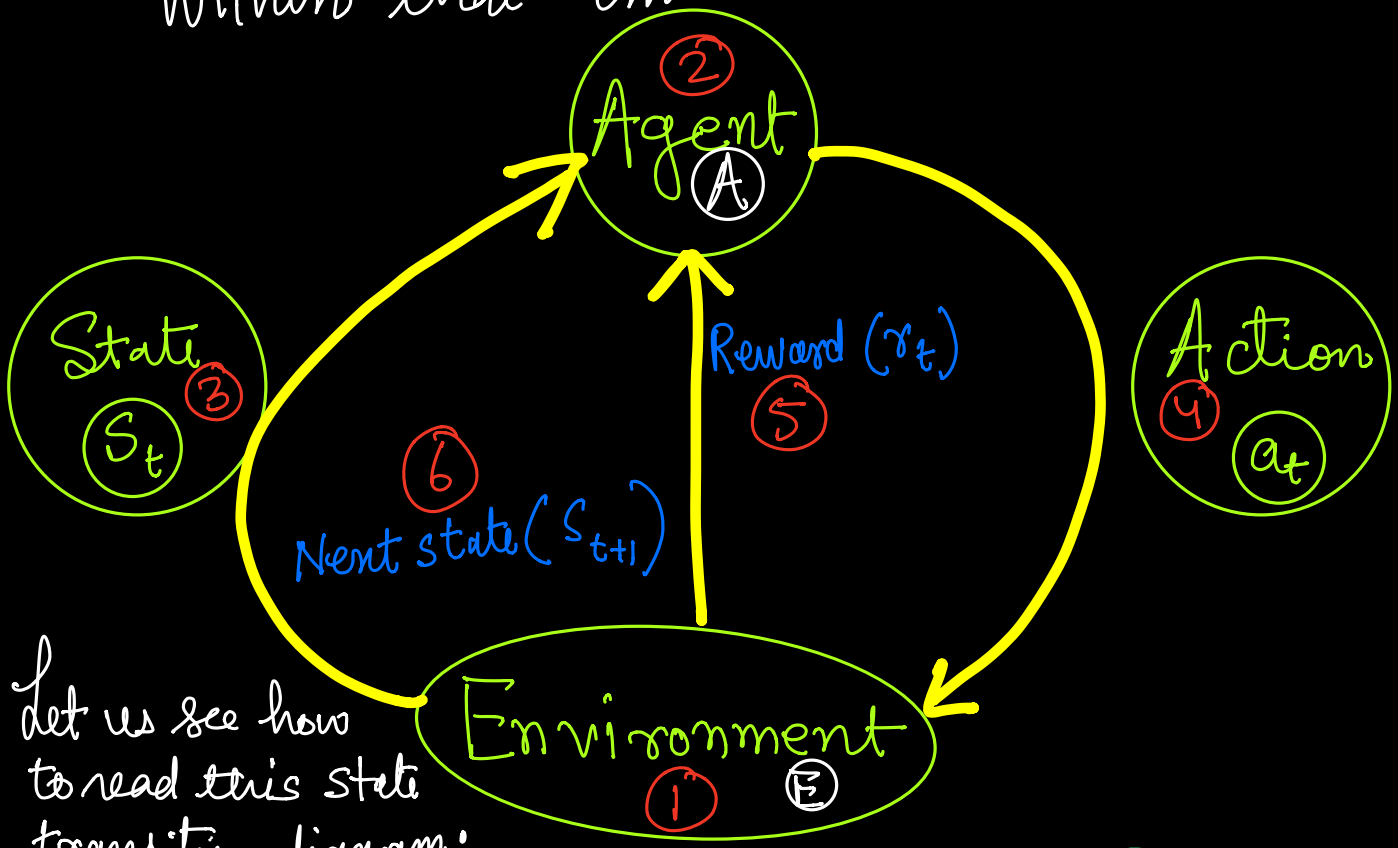
Examples: Clustering, dimensionality reduction, feature learning, density estimation.

(In the absence of target underlying data probability distributions can be estimated and used for various purposes. Such Generative models.)

(III) Reinforcement learning : It is a learning paradigm in which

learning happens by exploration. So agent interact with the environment repeatedly w/o any prior knowledge and labels wrt the enviromt.

— Completely relying on hit and trial strategy to learn how to behave optimally within that environment.



Let us see how to read this state transition diagram:

(*) The agent (A) is at some state (S_t) at time (t).

↳ It is interacting with Environment (E), via. few defined Set of actions (A).

Action at any state is choosen only by utilizing the State (S_t). It assumes that (S_t) Completely Characterises the Environmental state. (Req. No History) Markov

(*) At any state (s_t), taking a particular action say (a_t) is going to fetch a reward (r_t).

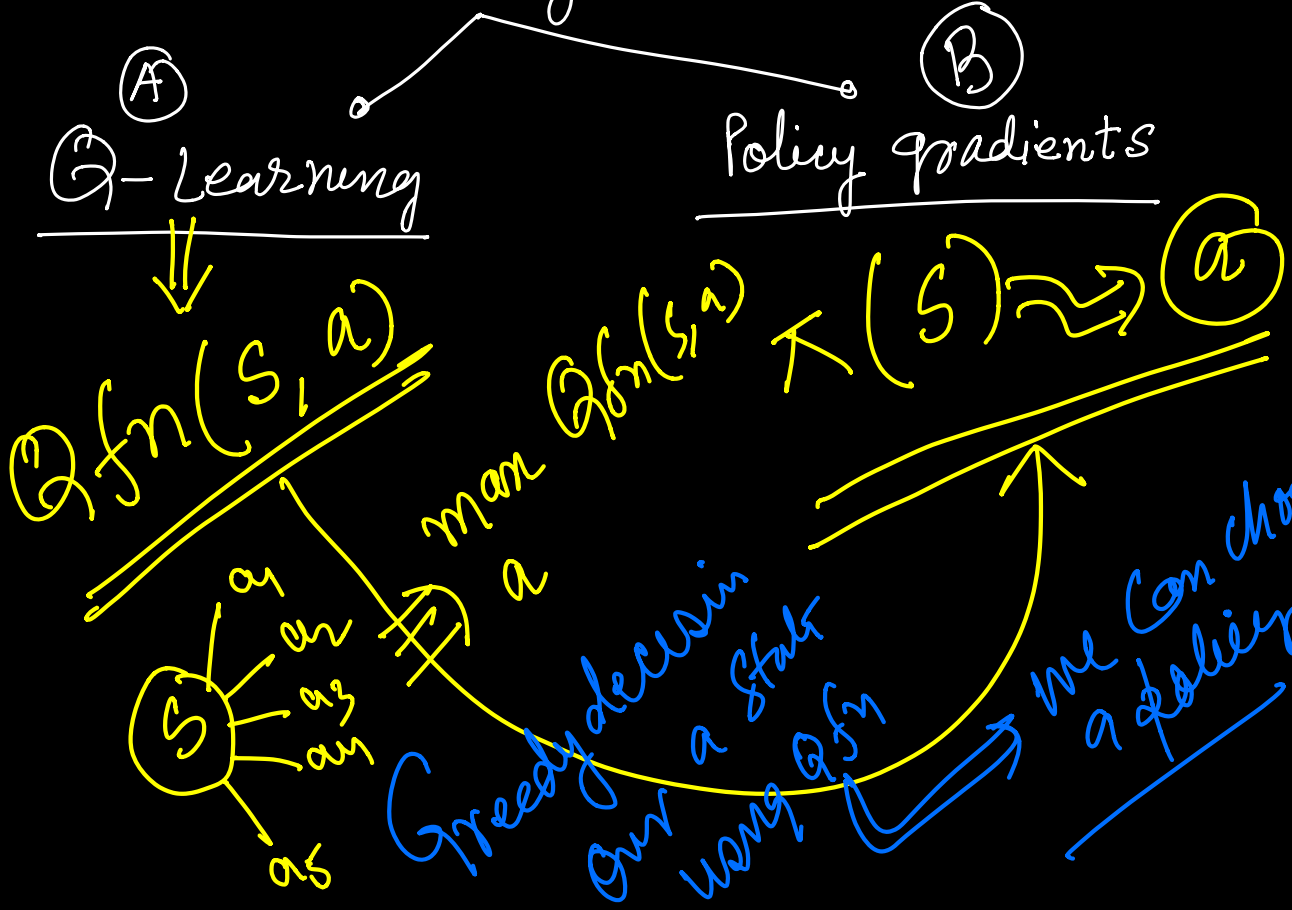
(*) The final Goal of the agent is :

To learn how to take an appropriate (optimal) action so as to maximize the reward.

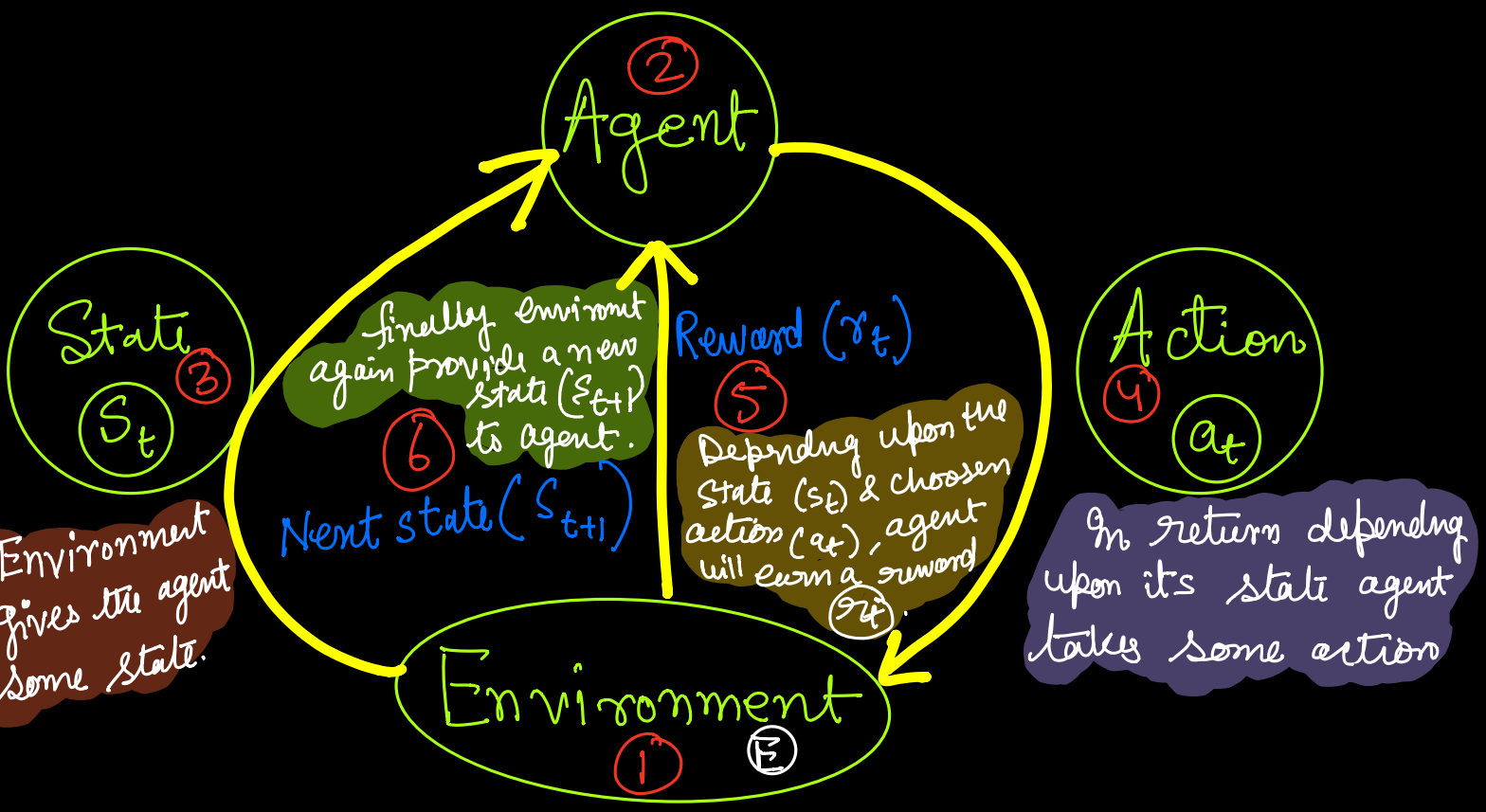
(Future discounted reward maximization)

(*) Re-inforcement learning can be formulated as Markov decision process $\Rightarrow$ MDP.

(*) There are (2) major classes of RL- algorithms



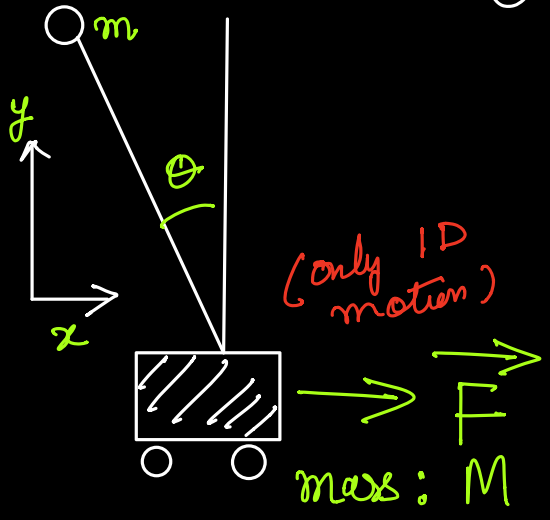
(B) → Basic RL-setup/framework



Episodic training : These six steps keeps on repeating in a loop and considered as an episode. The episode continues until the environment provides a terminal state (final) to the RL-Agent, after which the current episode get terminated.

(*) Hence one need to generate data (episodes) by allowing agent to explore the environment in episodes. Later use these episodic data for training the agent to learn how to behave optimally wrt the environment.

Example-01 : Cart Pole problem.



(only 1D motion)

(A) Objective : To balance the pole on the top of a movable Cart.

(B) State : The physical state can be encoded as : angle (θ), angular speed (ω), position (x, y), horizontal velocity (v_x) (may be more)

(C) Action : Horizontal force vector applied on the Cart.

(D) Reward : (+1) each time, if the pole is upright.

Example-02 : Atari game playing agent.

(A) Objective : Finish the game with the highest score.

(B) State : Game state can be raw pixel input of the game screens. (Image of game screen)

(C) Action : Game controlling actions, (L / R / U / D).

(D) Reward : At the end of the episode if win then all <state, action> pair will get a (+1) reward else they will get (−1) reward.

C → Mathematical formulation of RL.

(*) RL can be mathematically formulated using Markov Decision processes (MDP).

C.1 Markov property:

The future is independent of the past given the present.

Def: A state (S_t) is Markov iff

$$P[S_{t+1} | S_t] = P[S_{t+1} | S_1, S_2 \cdots S_t]$$

(*) State captures all relevant informations from the history. State is a sufficient statistic of the future.

C.2 Markov Decision Process:

MDP is a decision process, choosing action at each state. It is an environment in which all states are Markov. The MDP can be characterized by tuple:

$$\langle S, A, P, R, \gamma \rangle$$

- **S**: Set of all possible states: $s_1, s_2 \dots s_t \dots s_N$.
- **A**: Set of all possible actions: $a_1, a_2 \dots a_t \dots a_K$
- **P**: State transition Matrix: $\langle s_t, a_t \rangle \downarrow s_{t+1}$
- **R**: Reward fn: $\langle s_t, a_t \rangle \downarrow r_t$
- **γ**: Discount factor (0-1) (How much current rewards are important as compared to future.)

$\langle s_t, a_t \rangle$ → P → s_{t+1}, → R → r_t

C.2.1 State transition Matrix (P)

Markov process/chain is a memory less random process of sampling iteratively as per the given STM (P), starting from some given (seed) random starting state and following Markov property.

Full game dynamics can be encoded within (P). Robot can be interacting with with the gaming environment follows (P) $\Rightarrow$ The game rules.

$$P^{a}_{s\,s'} = P\left[S_{t+1} = s' \,\middle|\, S_t = s, a_t = a\right]$$

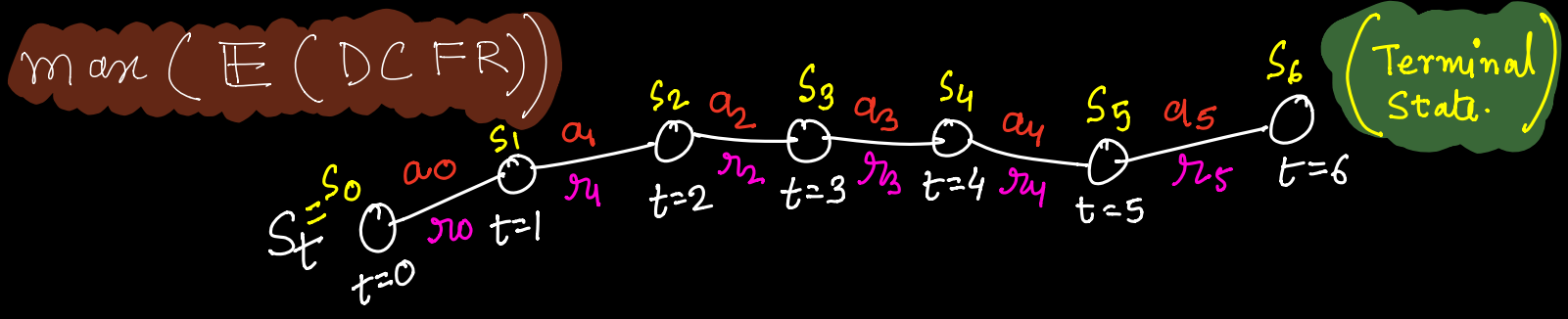
C.2.2 Immediate reward

Amount of reward an agent is going to get from the environment when at state S_t agent takes an action a_t.

$$r_{s_t}^{a_t} = r(s_t, a_t)$$

Greedy approach

* But in MDP, we are not intessted in immediate rewards. An optimal behaviour must need to maximize the Average Discounted Cumulative Future Reward (DCFR)



How MDP operates and can be used to obtain Episodes.

MDP- Execution

① At time ($t=0$), environment may sample randomly an initial state $s_0 \sim p(s_0)$

Initial state Probability distribution (ISPD)

② from $t=0$ (until terminates)

②.① An agent can select an action (a_t) [HOW!!!]

Using some policy that we need to figure out so as to max ($\mathbb{E}$(DCFR))

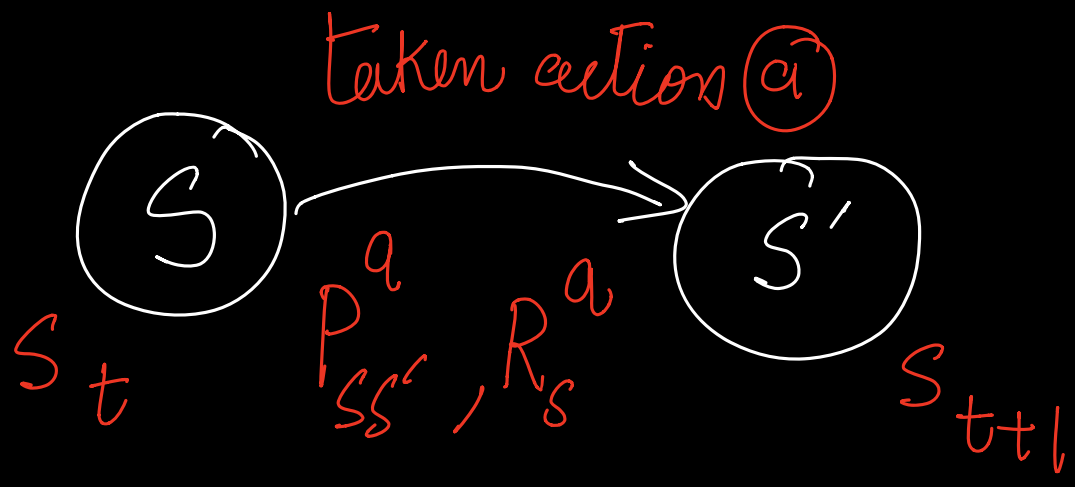
(2.2) Depending upon the agent's state S_t and the action taken a_t environment will return r_t the immediate reward as well as next state S_{t+1} using R & P.

∴ for a MDP P & R governs.

$$\Rightarrow P^{a}_{ss'} = P\left[S_{t+1} = s' \mid S_t = s, A_t = a\right]$$

$$\Rightarrow R^{a}_{s} = \mathbb{E}\left[r_t \mid S_t = s, A_t = a\right]$$

given a discount factor $\gamma \in [0,1]$



C.2.3 But how to take an action ⇒ Policy (π)

A policy (π) is just a mapping of states (S_t) to actions (a_t) enabling any agent to take action

$$\mathbb{R}^{\text{state}} \longrightarrow \mathbb{R}^{\text{action}}$$

(*) It can be deterministic or stochastic (Random). Defined

as $\pi(a_t | s_t) = P[a_t | s_t]$

C.2.4 Goal of MDP

The objective of MDP is to find an optimal policy (π^*), that can maximizes the DCFR.

(*) The actual returns are always in the form of DCFR.

G_t = Discounted cummulative future Reward from time step (t). [future rewards are geometrically weighted]

$= r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots$

for all upcoming future time step

$$= \sum_{K=0}^{\infty} \gamma^K R_{t+K} \qquad (\gamma \in (0,1))$$

{Gamma}

γ → close to (0) "Myopic" evaluation.

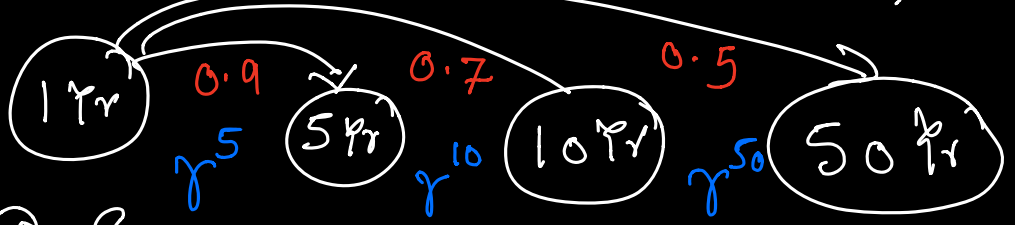
More importance to immediate returns.

→ close to (1) "far-sighted"

Care about all future rewards

(*) Requirement of discount :

(a) future is uncertain, and we care about :
If I invest 1K-INR (Today) ⟹ then how do I need to care about my return after (1 Yr), (5 Yr), (10 Yr) --- .



$\gamma \in \{0-1\}$

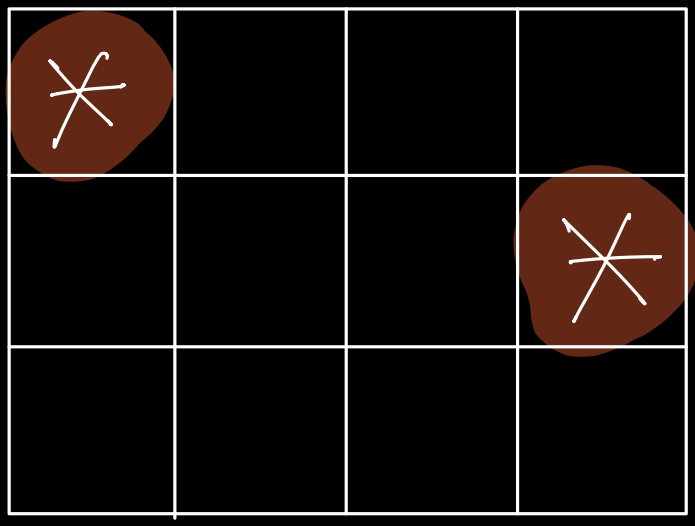
(b) Since we are predicting using an inaccurate model of environment hence the trust over model predictions can decrease exponentially. [Reason]

(c) Mathematically Convenient to discount rewards.

(d) Avoids infinite returns in Cyclic Markov processes.

C.2.5 Example Grid World – MDP

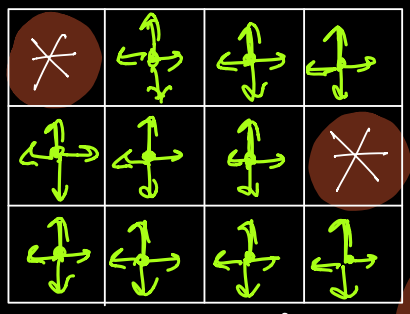
(A) States (2D-locations) (B) Actions = {right, left, up, down}.



(C) Reward = $\{r_t = -1\}$ negative reward for each transition. It ensures to learn how to reach target fast.

(D) Objective: Starting from any randomly choosen block ⟹ Reach to a terminal state (*) in the least number of actions. !

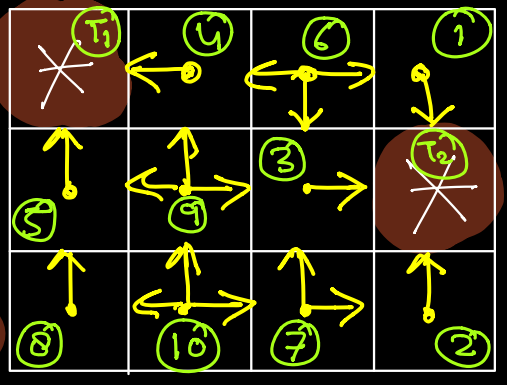
(E) Given the above 2D-Grid world MDP we need to figure out some optimal policy that can maximize the DCFR. (Out of all the policies π)



Random policy (A) (π_1)

Randomly choosing any direction.

Visual policy realization



(B) (π_2)

Policy philosophy: (a) For immediate neighbours:

One should have to move in a direction that can take you to the terminal state.

(b) Other states: Move in the direction that can take you closest to one of the immediate neighbouring state. $\pi(a_t|S_t)$ [Mapping from $S_t \to a_t$]

(D) → Solution of MDP: Optimal Policy (π^*)

Policy ≋> Planning

It do not matter that in whatever state (say S_t) agent is at time t, (π^*) is going to suggest that "what actions (a_t) agent needs to take in order to get the DCFR Maximized. $\forall_{a_i} \pi(a_i|S_t)$

(*) Such MDP's are realized as stochastic framework with initial probability distribution or transition probability distribution. Hence in order to handle the randomness we always talk about Expected value of the reward maximization.

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k * r_{t+k} \,\middle|\, \pi\right]$$

Returning its best policy

Maximizing — DCFR

Expectation over all the policies

Given a policy.

where $s_0 \sim p(s_0)$

$a_t \sim \pi(\cdot \mid s_t)$

$s_{t+1} \sim \mathbb{P}(\cdot \mid \langle s_t, a_t \rangle)$

The initial state, action & next state got sampled from their respective probability distribution.

(*) There are few attributes associated to a policy:

$\pi(a_t \mid s_t)$

(A) Value fn — wrt a given policy (π), How good is a given state (S), i.e if agent just follow (π) How much DCFR is expected is the value of that state (S) for a given policy (π).

(B) Q-value fn — wrt a given policy (π), How good is a state action pair. i.e Instead of directly following the policy (π) at state (S), agent first takes an action (a) and then follow the policy (π). Q value of $\langle s, a \rangle$ is the expected DCFR.

D.1 → Value function $[V^{\pi}(s)]$

Defined as the expected DCFR if agent follow π at state s until the episode terminates (reaching terminal state)

↳ Given a policy and a initial state $s_0 \sim p(s_0)$; Episode is $(s_0, a_0, r_0) \rightarrow (s_1, a_1, r_1)$

Episode carry on till terminal state is reached <--- $\leftarrow (s_3, a_3, r_3) \leftarrow (s_2, a_2, r_2)$

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \,\middle|\, S_t = s, \pi\right]$$

Value of state s wrt a given policy π

Expected DCFR from state s following the policy π to choose future actions.

D.2 → Q-value function $[Q^{\pi}(s,a)]$

Let us assume that we have

$S = \{s_1, s_2, s_3\}$ $A = \{a_1, a_2\}$

for some (π) and state (s_2)

① $V^{\pi}(s_2)$

we can have ⑥ state action pairs

s_1, a_1	s_2, a_1	s_3, a_1
s_1, a_2	s_2, a_2	s_3, a_2

② $\frac{Q^{\pi}(s_2, a_1)}{Q^{\pi}(s_2, a_2)}$

③ $\pi(a_1|s_2)$ & $\pi(a_2|s_2)$

④ $\pi^*(a_1|s_2)$ & $\pi^*(a_2|s_2)$

$Q^{\pi}(s,a)$ ⟹ How good it is to take an action (a) at state (s) & then following (π).

Basically ⟸ the expected DCFR, if at state (s), (a) action has been choosen.

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^{k} r_{t+k} \,\middle|\, s_t = s,\ a_t = a, \pi\right]$$

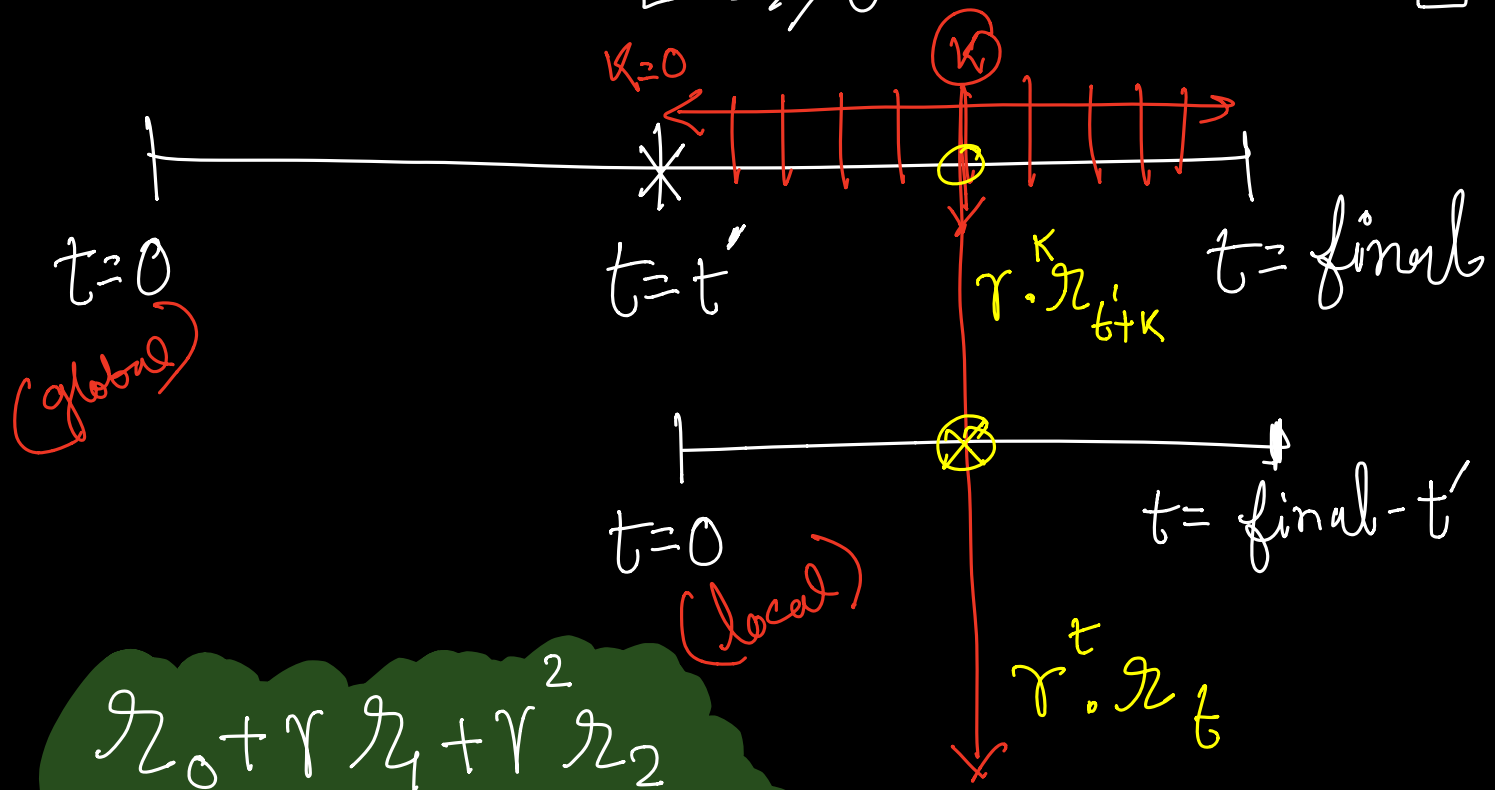
DCFR

Used for policy optimization

at state (s) agent took action (a) then follow (π)

Equivalently

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t \;\middle|\; s_0 = s,\ a_0 = a,\ \pi\right]$$



$$r_0 + \gamma r_1 + \gamma^2 r_2 + \gamma^3 r_3 + - - - - -$$

0.3 → Optimal value function and Q-value fn.

(*) Optimal state value fn $V^*(s)$

↳ maximum value function over all the policies

$$V^*(s) = \max_{\pi} V_{\pi}(s)$$

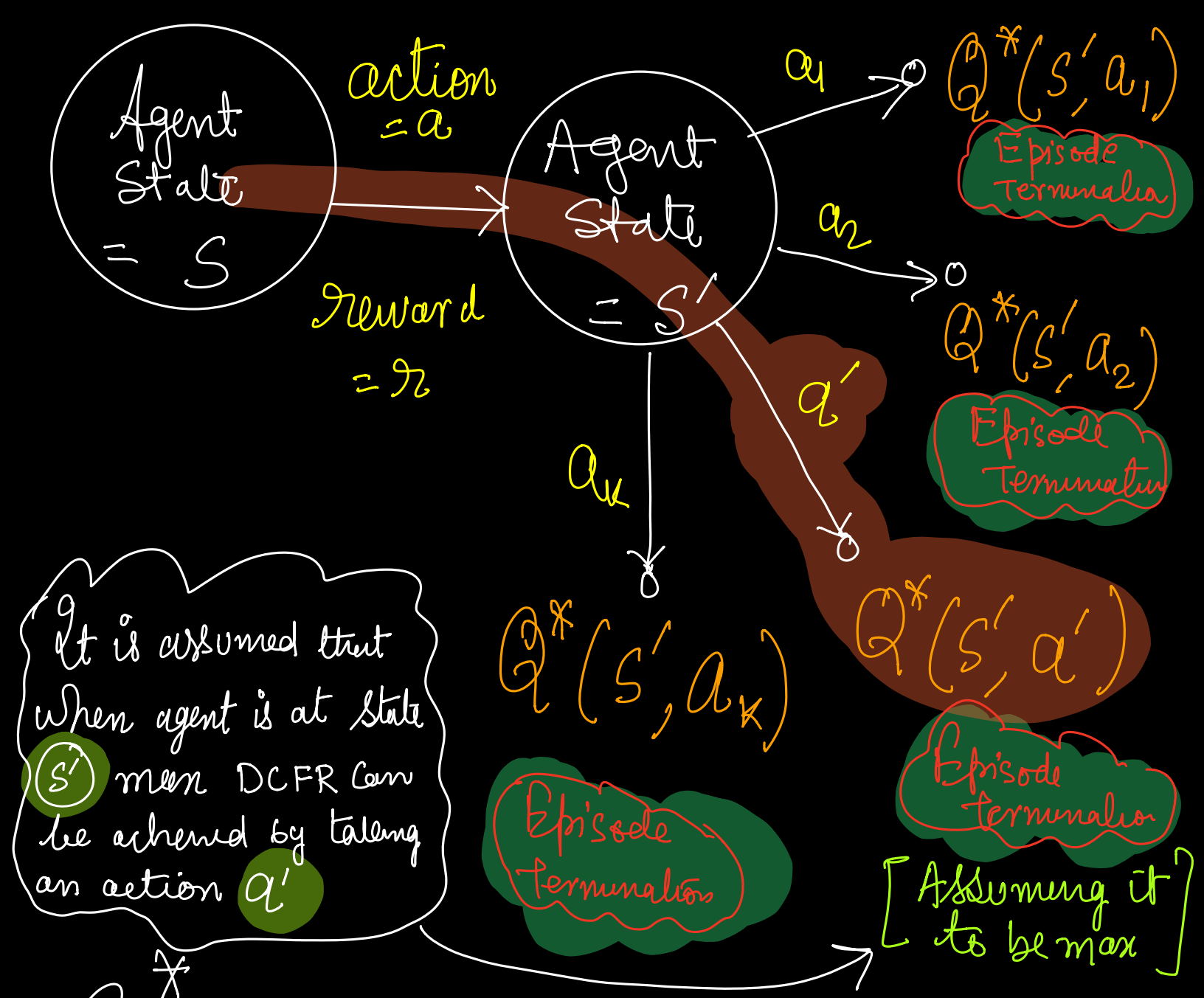
(*) Optimal action-value of Q-value fn $Q^*(s,a)$

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a)$$

$$= \max_{\pi} \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t \,\middle|\, \begin{matrix} s_0 = s \\ a_0 = a \\ \pi \end{matrix} \right]$$

(Maximizing over all policies)

Maximum DCFR one can achieve from any given (state, action) pair.



It is assumed that when agent is at state S' then DCFR can be achieved by taking an action a'

[Assuming it to be max]

$$Q^*(s,a) = r + \gamma * \max(Q^*(s', a_1), Q^*(s', a_2), \dots, Q^*(s', a_k))$$

[best Q-value over all actions] $\Leftarrow$

$$= r + \gamma * Q^*(s', a')$$

Since at state (s) after taken an action (a) agent will land into state (s') with some probability (p).

$\Rightarrow$ $s, a, s_1 = 0.3$; $s, a, s_2 = 0.4$; $s, a, s_3 = 0.3$

$\therefore Q^*(s,a) = \mathbb{E}_{s' \in \epsilon}\left[r + \gamma \max_{a'} Q^*(s', a') \mid s, a \right]$

it may reach to a set of states (ϵ)

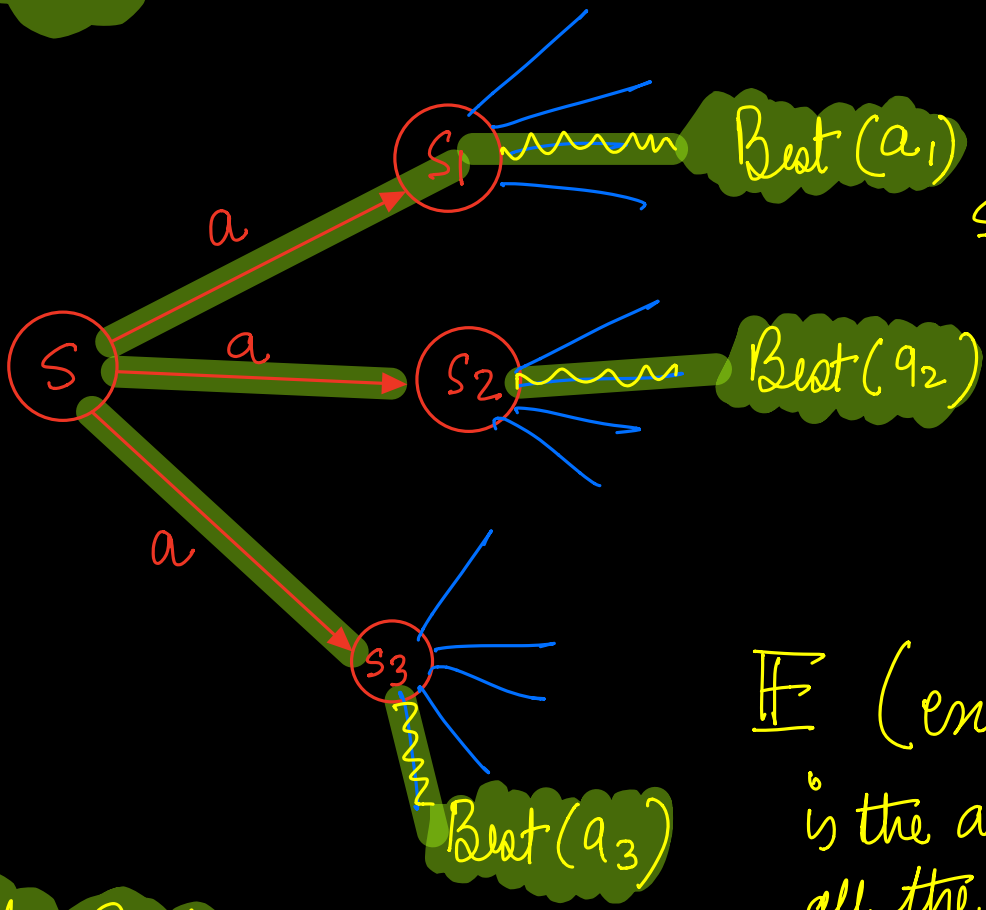
Q^* as well as V^* both functions satisfies Bellman equation and are defined recursively.

(*) Expected values are considered in order to address the randomness in the process.

Just see it once more as we are going to use it for Deep Q Learning (Q-fn = Quality of $\langle s, a \rangle$ wrt π)

$$Q^*(s,a) = \mathbb{E}_{s' \in} \left[r + \gamma \max_{a'} Q^*(s', a') \mid s, a \right]$$

it may reach to a set of states ($\in$)



$s' \Rightarrow (s_1, s_2, s_3)$

$\in = \{s_1, s_2, s_3\}$

$a' =$ ← as shown

$\mathbb{E}$ (expected value) is the average of all the (3)

basically whatever we observe after taking a decision (a) at state (s)

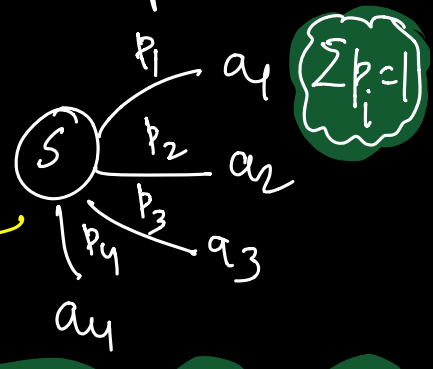
D.4 → Optimal Policy (π^*)

Policy: Def:

(*) MDP policies depends only over Current state. (Not on History)

(*) Hence policies are stationary (Time-independent).

$$\pi(a|s) = \mathbb{P}[A_t = a \mid S_t = s]$$



Here p_i's we are deciding under policy assumg to be given. Later we need to Compute these optimally using DP ⇒ Not Scalable hence approximate it using Policy Network.

Stochastic policies ensures suitable state, action space exploration

(*) Policy governs the dynamics of agent.

(*) Best/optimal policy need to ensure maximum future reward.

(*) In Markov Random Processes, we don't care about the historical reward accumulated so far. We only Considered DCFR (Discounted Cumulative future reward.)

⇒ This is how we behave optimally

(E) → How to solve for optimal policy (using DP)

(*) firstly we need to figure out how to evaluate a policy (Policy Evaluation) This will tell us how good is that policy.

(*) Later using Bellman equation policy iteration or value iterations can be done inorder to update the current policy. (Iterative update)

(*) Basically assuming policy or value function at (i^{th}) step are optimal and will be used to compute the policy/value for $(i+1^{th})$ step. [Using Dynamic Programming]

E.1 → Policy Evaluation

1. $\pi \equiv$ Policy under Evaluation

State to action mapping as per policy

4

$\pi(a_1|s)$

5

2

S

→ Computing $V_{K+1}(s)$!!!

using

V_K

3

$\pi(a_2|s)$

6

a_1 7

a_2 8

$R_s^{a_1}$ 9

$R_s^{a_2}$ 10

Rewards as per MDP

State transitions as per MDP dynamics

11 (a) $P_{ss_1}^{a_1}$

12 (b) $P_{ss_2}^{a_1}$

13 (c) $P_{ss_3}^{a_2}$

14 (d) $P_{ss_4}^{a_2}$

$R_{ss_1}^{a_1}$ $R_{ss_2}^{a_1}$ $R_{ss_3}^{a_2}$ $R_{ss_4}^{a_2}$

S_1 15 S_2 16 S_3 17 S_4 18

$V_K(S_1)$ $V_K(S_2)$ $V_K(S_3)$ $V_K(S_4)$

(a)

$$\left[R_{ss_1}^{a_1} + P_{ss_1}^{a_1} * V_K(S_1)\right]\pi(a_1|s)$$

(c)

$$\left[R_{ss_3}^{a_2} + P_{ss_3}^{a_2} * V_K(S_3)\right]\pi(a_2|s)$$

$$\left[R_{ss_4}^{a_2} + P_{ss_4}^{a_2} * V_K(S_4)\right]\pi(a_2|s)$$

(d)

(b)

$$\left[R_{ss_2}^{a_1} + P_{ss_2}^{a_1} * V_K(S_2)\right] * \pi(a_1|s)$$

Assuming 15, 16, 17, 18 is given/ known/ initialized

(Their discounted values are added)

$$V_{K+1}(s) = a + b + C + d$$

(Expected DCFR of (S) following π and using $V_K^{'s}$

$$= \sum_{a \in A} \pi(a|s) * \left[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a * V_K(s') \right]$$

This probability is doing the expectation.

In closed form:

$$V_{K+1} = R^{\pi} + \gamma P^{\pi} V_K$$

- V_{K+1}: Value vector for all states at $(K+1)^{th}$ iteration
- R^{π}: Reward function as per the MDP
- P^{π}: State Transition Probability
- V_K: Value for at K^{th} iteration

Let us evaluate a random policy in small grid world.

E.2 Policy Evalution : Example

F.2 Deep Q-learning for Q fn approximation

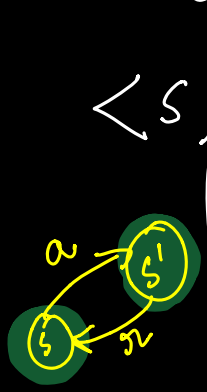
* Instead of computing $Q(s, a)$, we need to use some function approximator (say neural network) to approximate/estimate $\{Q(s, a) \; \forall \; s, a\}$

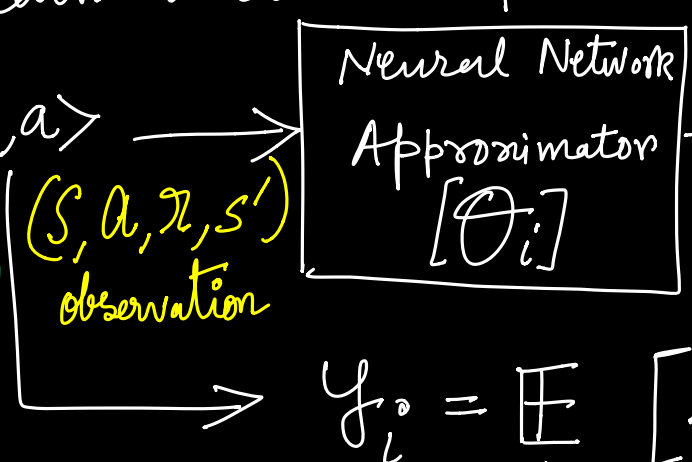
$$Q(s, a; \theta) \rightsquigarrow Q^*(s, a)$$

Using DNN parameters to estimate the Q value for (s, a) pair. { Deep Q-Learning }

* We need a Q fn approximator that satisfies Bellman equation by enforcing Bellman optimality at each iterative step.

(Intuition)





$Q(s, a; \theta_i)$

$Q_t(s, a; \theta_t)$

$$y_i = \mathbb{E}_{s' \sim \epsilon}\left[r + \gamma \max_{a'} Q(s', a'; \theta_i \mid s, a)\right]$$

$Q(s, a; \theta_{t-1})$

(6) Solving for Optimal Policy (Directly)

6.1 Introduction to policy gradient

(*) Instead of using DNN to approximate Q-value function, why can't directly learn the suitable policy (π), parameterized by θ.

Class of parameterized policies : $\pi = \{\pi_\theta \mid \theta \in \mathbb{R}^m\}$

for any policy (π), define it value as : $J(\theta) = \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t \mid \pi_\theta\right]$

- $J(\theta)$: value of the policy π parameterized by θ
- $\mathbb{E}$: Expected
- $\sum_{t \geq 0} \gamma^t r_t$: DCFR
- π_θ : while following policy π_θ

(*) Given any policy (π_θ), how much DCFR on an average one can extract $\Rightarrow$ value of π_θ i.e. $J(\theta)$

GOAL : DNN need to optimize for the parameter θ that can realize the optimal policy π^*.

$\theta^* = \arg\max_\theta J(\theta)$ Gradient Ascent over θ $\rightsquigarrow$ REINFORCE also.

(A) Deep Q-Learning

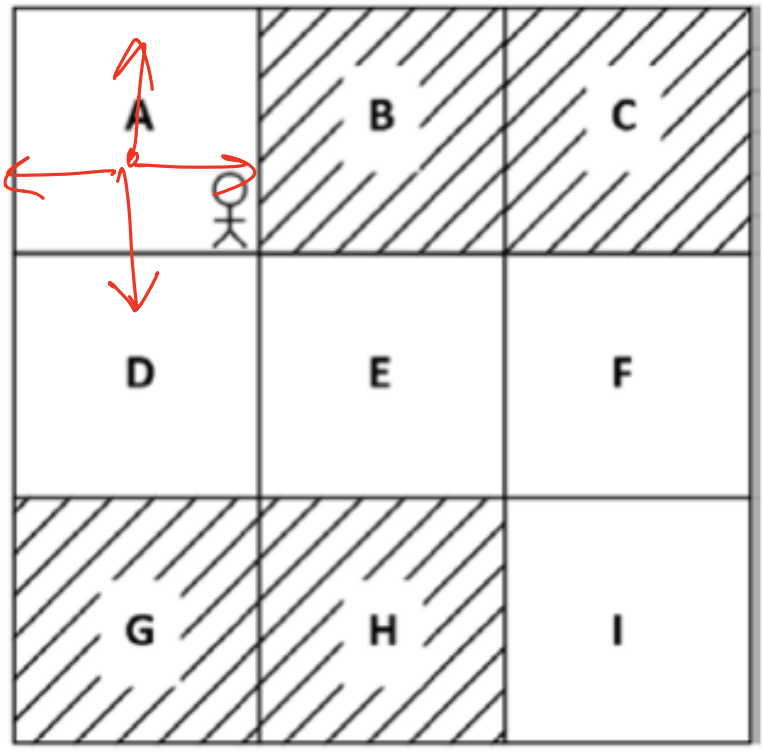


Figure 9.1: Grid world environment

State	Action	Value
A	up	17
A	down	10
B	up	11
B	down	20

Table 9.1: Q-value of state-action pairs

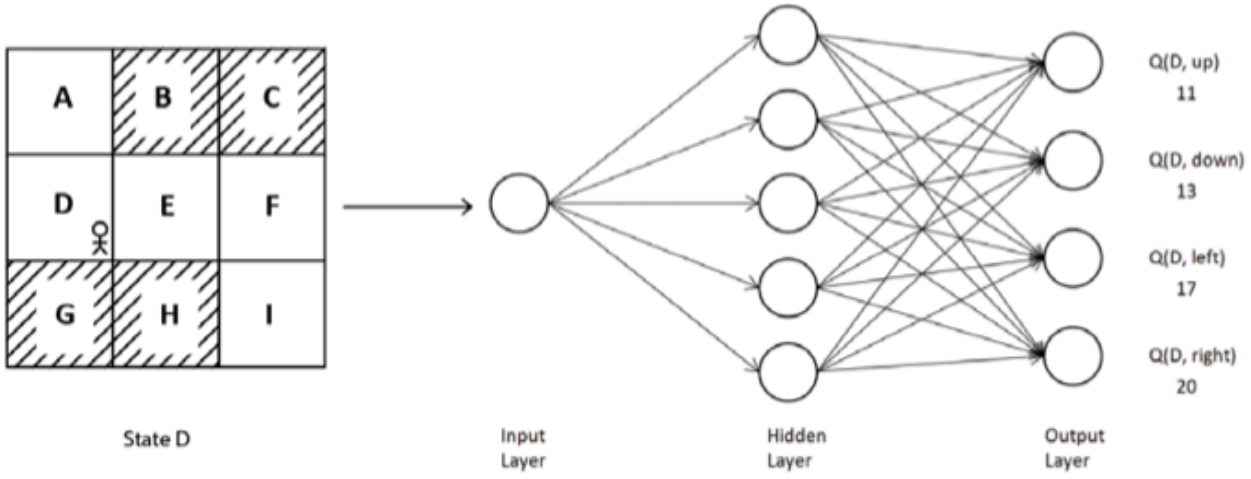


Figure 9.2: Deep Q network

DQN - Learning Q fn.

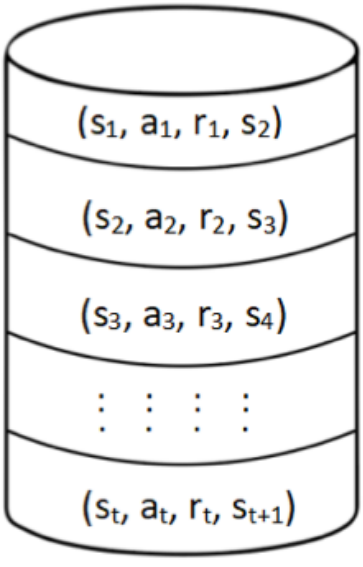


Figure 9.3: Replay buffer

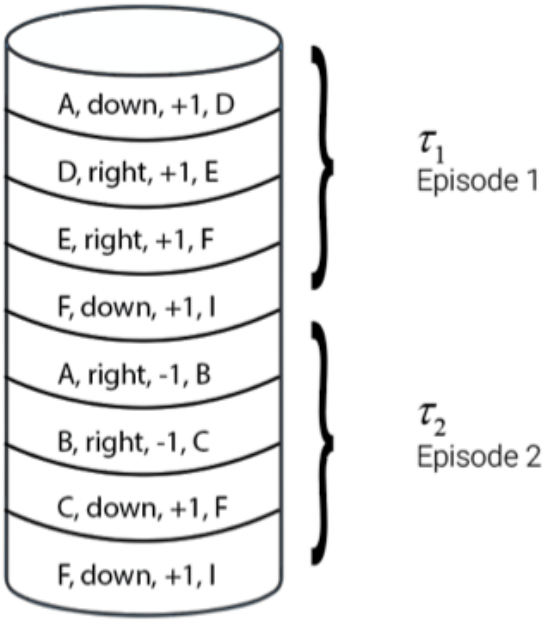


Figure 9.6: Replay buffer

the replay buffer $\mathcal{D}$.

1. Initialize the replay buffer $\mathcal{D}$.
2. For each episode perform *step* 3.
3. For each step in the episode:
    1. Make a transition, that is, perform an action a in the state s, move to the next state s', and receive the reward r.
    2. Store the transition information (s, a, r, s') in the replay buffer $\mathcal{D}$.

1

$$\text{MSE} = \frac{1}{K}\sum_{i=1}^{K}(y_i - \hat{y}_i)^2$$

2

$$Q^*(s,a) = \mathbb{E}_{s' \sim P}[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')]$$

3

$$Q^*(s,a) = \underset{s' \sim P}{E}\left[r + \gamma \max_{a'} Q^*(s',a')\right]$$

4

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

5

$$L(\theta) = Q^*(s,a) - Q_\theta(s,a)$$

6

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

7

$$L(\theta) = r + \gamma \max_{a'} \boxed{Q(s', a')} - Q_\theta(s, a)$$

How do we compute this?

8

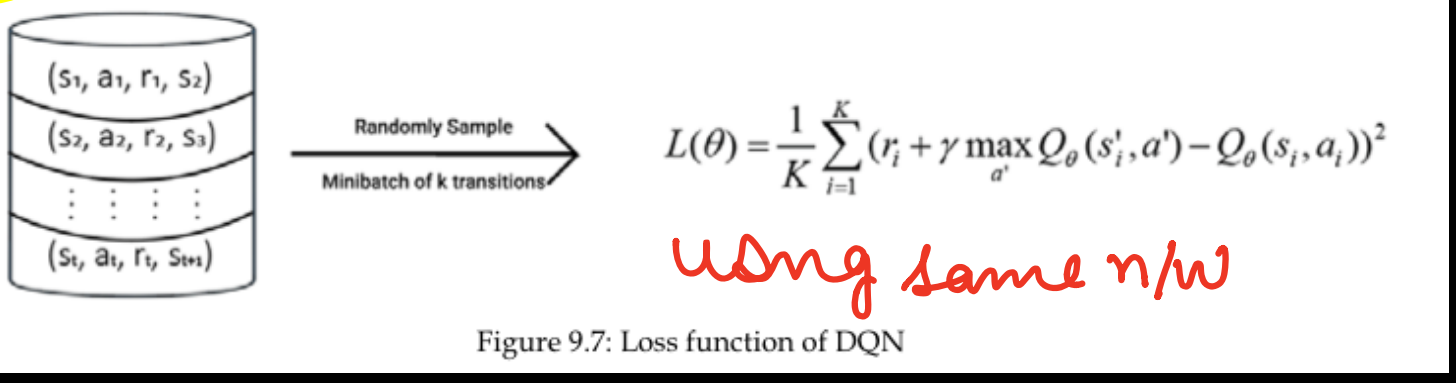


Figure 9.7: Loss function of DQN

9 Target N/W (Same n/w may introduce instability)

$$L(\theta) = \frac{1}{K}\sum_{i=1}^{K}(r_i + \gamma \max_{a'} \underbrace{Q_{\theta'}(s_i', a')}_{\text{Compute using } \theta'} - \underbrace{Q_\theta(s_i, a_i)}_{\text{Compute using } \theta})^2$$

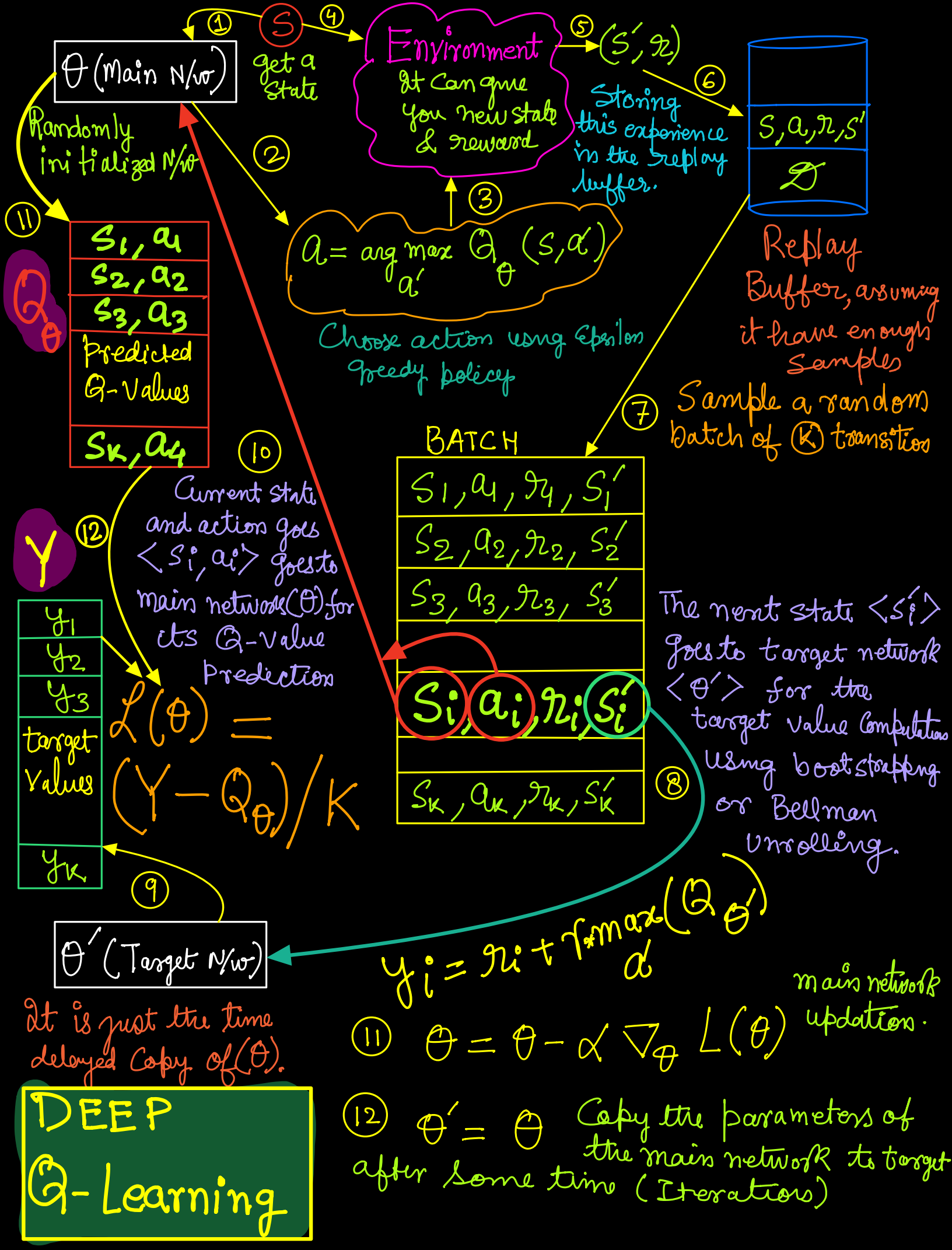
in the next section.

Final Algorithm:

The DQN algorithm

The DQN algorithm is given in the following steps:

1. Initialize the main network parameter θ with random values
2. Initialize the target network parameter θ' by copying the main network parameter θ
3. Initialize the replay buffer $\mathcal{D}$
4. For N number of episodes, perform *step 5*
5. For each step in the episode, that is, for $t = 0, \ldots, T\text{-}1$:
    1. Observe the state s and select an action using the epsilon-greedy policy, that is, with probability epsilon, select random action a and with probability 1-epsilon, select the action $a = \arg\max_a Q_\theta(s, a)$
    2. Perform the selected action and move to the next state s' and obtain the reward r
    3. Store the transition information in the replay buffer $\mathcal{D}$
    4. Randomly sample a minibatch of K transitions from the replay buffer $\mathcal{D}$
    5. Compute the target value, that is, $y_i = r_i + \gamma \max_{a'} Q_{\theta'}(s_i', a')$
    6. Compute the loss, $L(\theta) = \frac{1}{K}\sum_{i=1}^{K}(y_i - Q_\theta(s_i, a_i))^2$
    7. Compute the gradients of the loss and update the main network parameter θ using gradient descent: $\theta = \theta - \alpha\nabla_\theta L(\theta)$
    8. Freeze the target network parameter θ' for several time steps and then update it by just copying the main network parameter θ


θ (Main N/w)
Randomly initialized N/w
get a state
S
Environment
It can give you new state & reward
(S', r)
Storing this experience in the replay buffer.
S, a, r, S'
Replay Buffer, assuming it have enough samples
a = arg max Q_θ (s, a')
a'
Choose action using Epsilon greedy policy
Sample a random batch of (K) transitions
BATCH
S1, a1, r1, S1'
S2, a2, r2, S2'
S3, a3, r3, S3'
Si, ai, ri, Si'
SK, aK, rK, SK'
Q_θ
Predicted Q-Values
S1, a1
S2, a2
S3, a3
SK, a4
Current state and action goes <Si, ai> goes to main network(θ) for its Q-value Prediction
Y
y1
y2
y3
target Values
yK
L(θ) = (Y − Q_θ)/K
The next state <Si'> goes to target network <θ'> for the target value Computation using bootstrapping or Bellman unrolling.
θ' (Target N/w)
It is just the time delayed Copy of (θ).
y_i = r_i + γ max_a' (Q_θ')
main network updation.
⑪ θ = θ − α ∇_θ L(θ)
⑫ θ' = θ Copy the parameters of the main network to target after some time (Iterations)
DEEP Q-Learning

(B) Policy Gradient

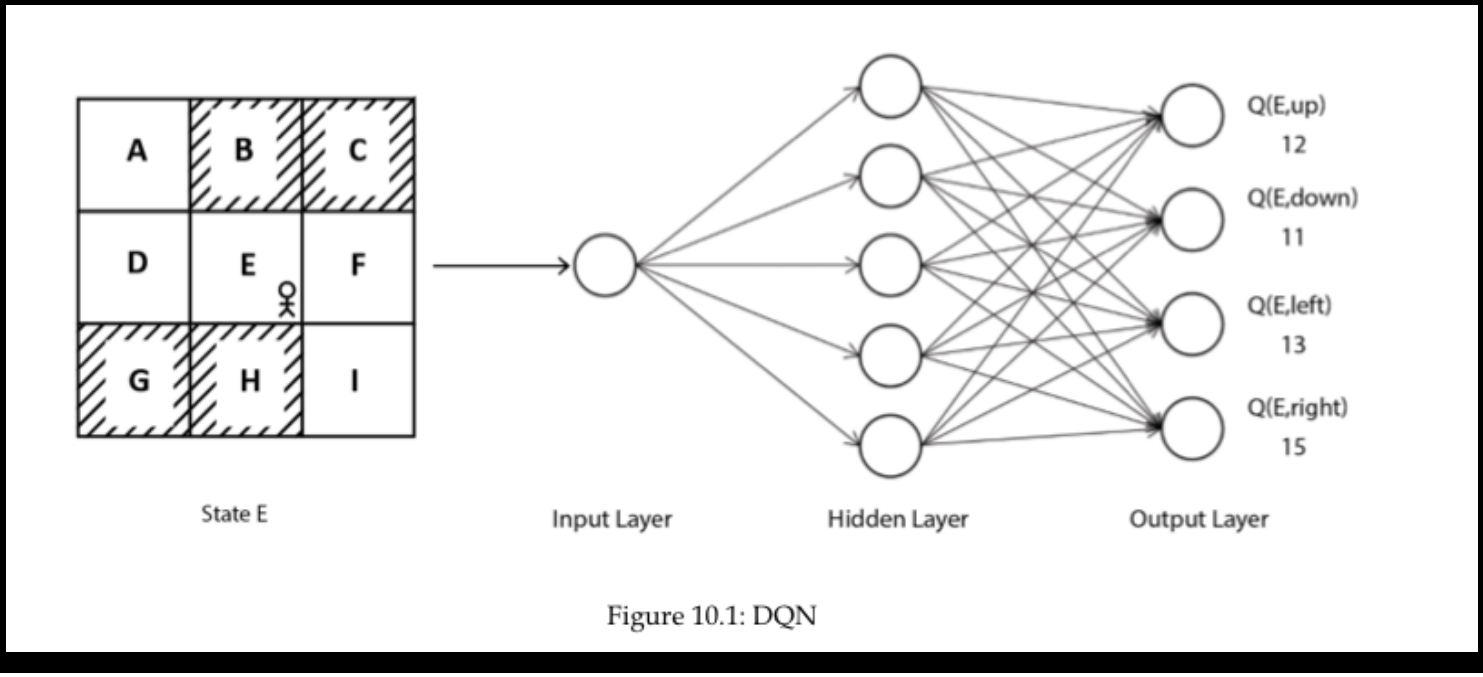


Figure 10.1: DQN

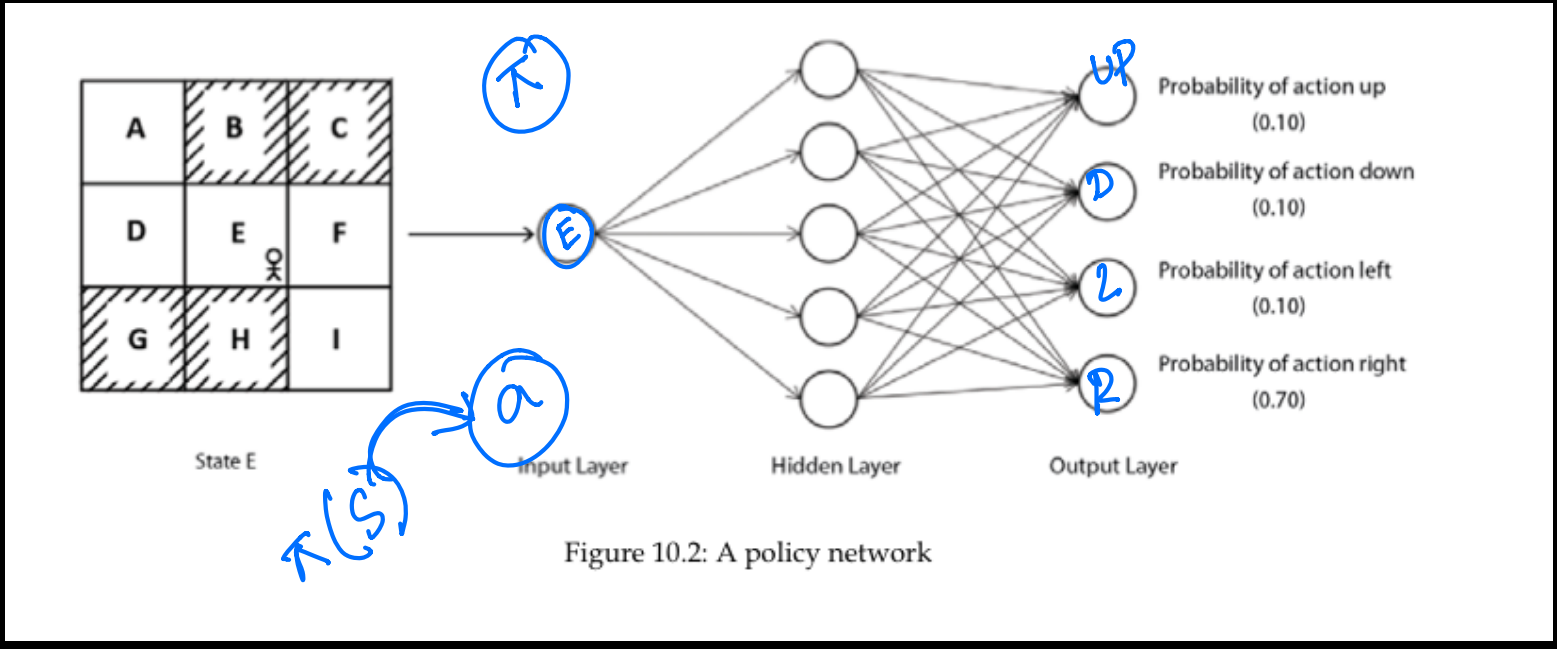


Figure 10.2: A policy network

In RL- we need to learn the optimal policy

(*) In Continous action space (like Car driving, Robot movement) DQN faces issues.

(*) Q fn parameterization (θ)

(*) Policy parameterization (θ)

$\Rightarrow$ we directly learn the policy.

Figure 10.3: Trajectory τ_1

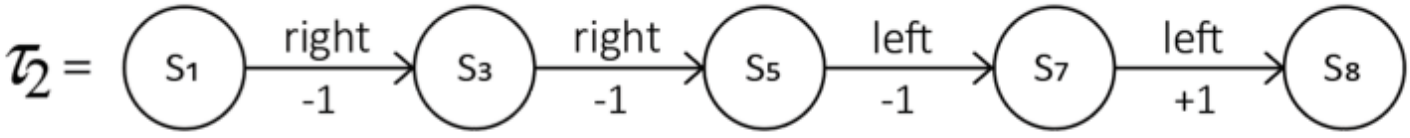


Figure 10.4: Trajectory τ_2

Intution:

(1) In PG, we use NN to approximate the optimal policy π^*.

(2) Initialize θ randomly.

(3) We feed state (S) as an input and it will return the action probabilities.

(4) Initially as N/w is not trained Random O/P action probabilities we get.

(5) Still we select the action based on the O/P action probability distribution.

(6) Also store (S, A, R, S') (untill the end of the episode).

(7) This will become our training data

(8) If agent WIN that episode ⇒ Assign high probability to all the actions of the episode for that (5) else reduce.

REINFORCE.

1. Initialize the network parameter θ with random values
2. Generate N trajectories $\{\tau^i\}_{i=1}^{N}$ following the policy π_θ
3. Compute the return of the trajectory $R(\tau)$
4. Compute the gradients

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)\, R(\tau)\right]$$

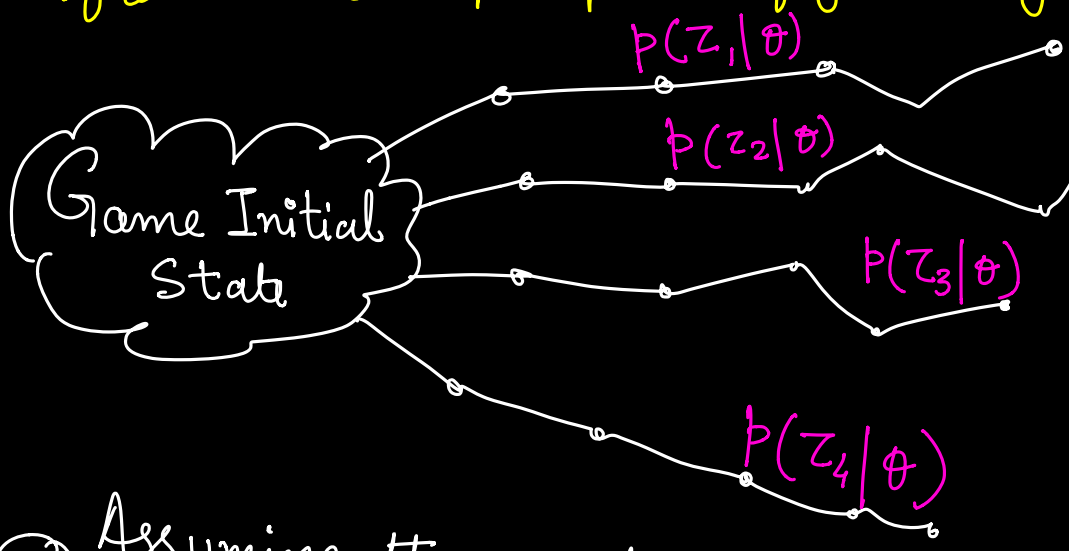
5. Update the network parameter as $\theta = \theta + \alpha\nabla_\theta J(\theta)$
6. Repeat *steps* 2 to 5 for several iterations

How to proof that $\nabla_\theta J(\theta) =$ as stated in step (4).

G.2 REINFORCE algorithm Loss function

(*) Trajectory (τ): $\tau = (<s_0, a_0, r_0>, <s_1, a_1, r_1>, <s_2, a_2, r_2>, - - - - - -)$

Sample trajectory from the example episode of game play. (Sampled randomly).



(*) Assuming the agent to be starting from some initial game state and following (π_θ) policy parameter over (θ).

→ (a) Assuming all trajectories to be equally probable: $J(\theta) = \frac{1}{4} \sum_{i=1}^{4} r(\tau_i)$

→ (b) Assuming trajectories follow $p(\tau|\theta)$ distribution

$$J(\theta) = \sum_{i=1}^{4} p(\tau_i|\theta) * r(\tau_i)$$

Value associated to a policy (π_θ), parameterized over [θ].

$$= \underset{\forall i}{\mathbb{E}} \left[r(\tau_i) \right]$$

Expected Reward $= \mathbb{E}_{\tau \sim p(\tau|\theta)}[r(\tau)]$

$$= \int_{\tau} (r(\tau) * p(\tau|\theta))\, d\tau$$

Considering all trajectories that one can encounter while agent start from some initial state & following (π_θ) policy.

$$\therefore \quad \theta^* = \arg\max_{\theta} [J(\theta)]$$

(*) REINFORCE algorithm uses $J(\theta)$ for its forward pass as a Gain fn [Need to be maximized]

G.3 REINFORCE algo Backward pass

(*) Forward Pass : $J(\theta) = \int_{\tau} r(\tau) * p(\tau|\theta)\, d\tau$ — (1)

(*) Backward pass: $\nabla_\theta J(\theta) = \int_{\tau} r(\tau) * \nabla_\theta p(\tau|\theta)\, d\tau$ — (2)

*** This is intractable, Gradient Computation of an expectation is problematic when p depends upon θ [Gradient of expectation]

Trick: $\nabla_\theta p(\tau|\theta) = p(\tau|\theta) * \left[\frac{\nabla_\theta p(\tau|\theta)}{p(\tau|\theta)}\right]$ — (3)

Putting eq (3) into eq (2) $= p(\tau|\theta) * \nabla_\theta \log[p(\tau|\theta)]$

$\therefore \nabla_\theta J(\theta) = \int_{\tau} r(\tau) * \nabla_\theta \log[p(\tau|\theta)] * p(\tau|\theta)\, d\tau$

$\therefore \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\tau|\theta)}\left[r(\tau) * \nabla_\theta \log[p(\tau|\theta)]\right]$

Expectations of Gradient

G.4 Issue with Gradient & its Computation

(*) There can be infinitly many trajectories possible for any parameterized policy (π_θ) estimated by DNN.

(*) The reward function [$J(\theta)$] has been defined as an Expectation over all such trajectories. Hence $J(\theta)$ Can be computed essentially by Monte-Carlo Sampling of trajectories.

(*) All trajectories ($\tau_1, \tau_2 \cdots \tau_\infty$) for some given ($\pi_\theta$) are not equiprobable, instead following a distribution. say $p(\tau|\theta)$

(*) Hence, for a some trajectory say [$\tau_1 = \langle (s_0, a_0, r_0), (s_1, a_1, r_1), (s_2, a_2, r_2) \rangle$]

Let us compute the probability of (τ_1) given $p(\tau|\theta)$

$$p(\tau_1|\theta) = \left[T(s_1|s_0, a_0) * \pi_\theta(a_0|s_0)\right] * \left[T(s_2|s_1, a_1) * \pi_\theta(a_1|s_1)\right] * \left[T(s_2|s_2, a_2) * \pi_\theta(a_2|s_2)\right]$$

Given the policy (π_θ) parameterized by (θ) & learned by some Deep policy N/w maximizing the reward fn [$J(\theta)$].

Transition Probabilities

As per Parameterized π_θ Policy.

(*) $$P(\tau|\theta) = \prod_{t \geq 0} T(s_{t+1}|s_t, a_t) * \pi_\theta(a_t|s_t)$$ Eq-A

Estimating the probability of randomly selected trajectory (τ), while following policy(π_θ)

The transition probability of going to (s_{t+1}) state after taken an action (a_t) at state (s_t)

Probability of taking an action (a_t) at state (s_t) under the policy (π_θ)

(*) The major problem is the $J(\theta)$ requires $P(\tau|\theta)$ which in turn requires transition probability (T) which is unknown. Our policy network is estimating (π_θ) by maximizing $J(\theta)$.

(*) Now the question is: Can we compute $J(\theta)$ without transition probabilities (T).

(*) But for backpropagation $\nabla_\theta[J(\theta)]$ only gradient of $J(\theta)$ is required and it is not depending upon (T).

As already shown

$$\therefore \nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p(\tau|\theta)} \left[R(\tau) * \nabla_\theta \log [p(\tau|\theta)] \right] \quad \text{(Eq- B)}$$

Expectation of Gradient

The log liklihood of such a sampled trajectory (τ) is using (Eq- A)

$$\log [p(\tau|\theta)] = \sum_{t \geq 0} \log T(s_{t+1}|s_t, a_t) + \log \pi_\theta(a_t|s_t) \quad \text{(Eq- C)}$$

Independent of [θ]

Dependent over [θ]

Just diffrentiating it wrt [θ].

$$\nabla_\theta \log [p(\tau|\theta)] = \sum_{t \geq 0} \nabla_\theta \log [\pi_\theta(a_t|s_t)]$$

Independent of transition probability

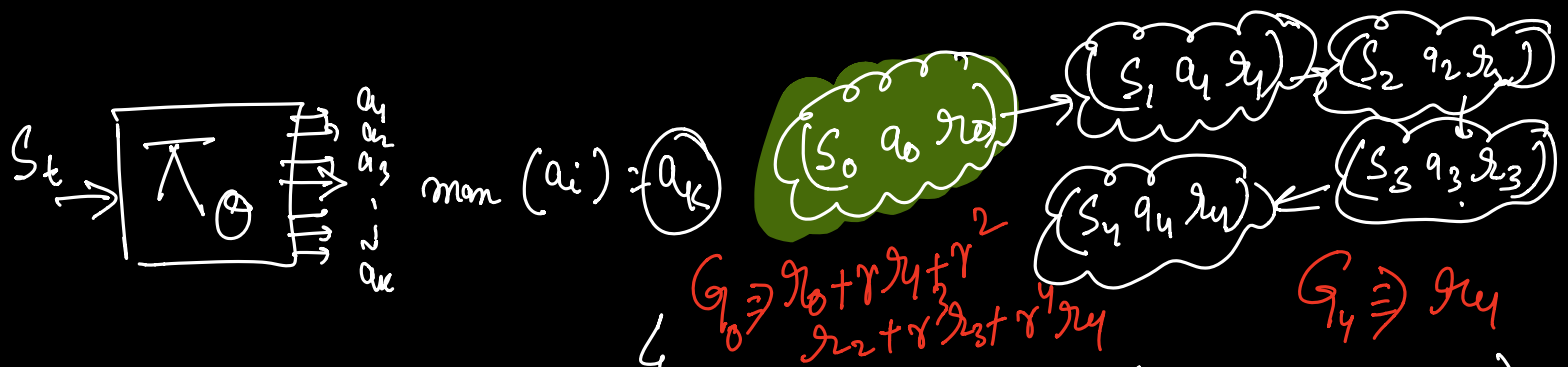
(Eq- D)

(*) Hence the derivate of log liklihood of the trajectory (τ), which is required for the computation of the Gradient of the Reward function i.e $\nabla_\theta (J(\theta))$ is independent of Transition Probability [T].

$\Longrightarrow$ This also implies that $\nabla_\theta (J(\theta))$ is also independent of [T]

Hence putting (Eq-D) into (Eq-B)

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim p(\tau|\theta)} \left[r(\tau) * \sum_{t \geq 0} \nabla_\theta \log \left[\pi_\theta(a_t|s_t) \right] \right]$$



$$\nabla_\theta (J(\theta)) = \sum_{t=0}^{4} \nabla_\theta \log \left(\pi_\theta (a_t|s_t) \right) * G_t$$

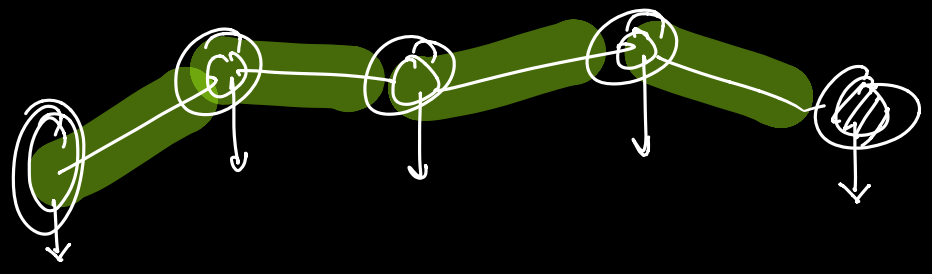
$$\Rightarrow G_0 * \nabla_\theta \log \left(\pi_\theta (a_0|s_0) \right) +$$
$$G_1 * \nabla_\theta \log \left(\pi_\theta (a_1|s_1) \right) +$$
$$G_2 * \nabla_\theta \log \left(\pi_\theta (a_2|s_2) \right) +$$
$$G_3 * \nabla_\theta \log \left(\pi_\theta (a_3|s_3) \right) +$$
$$G_4 * \nabla_\theta \log \left(\pi_\theta (a_4|s_4) \right)$$

$R(z)$ Good

$R(z)$ (Gradient Estimation)

Credit-Assignment

unbiased.



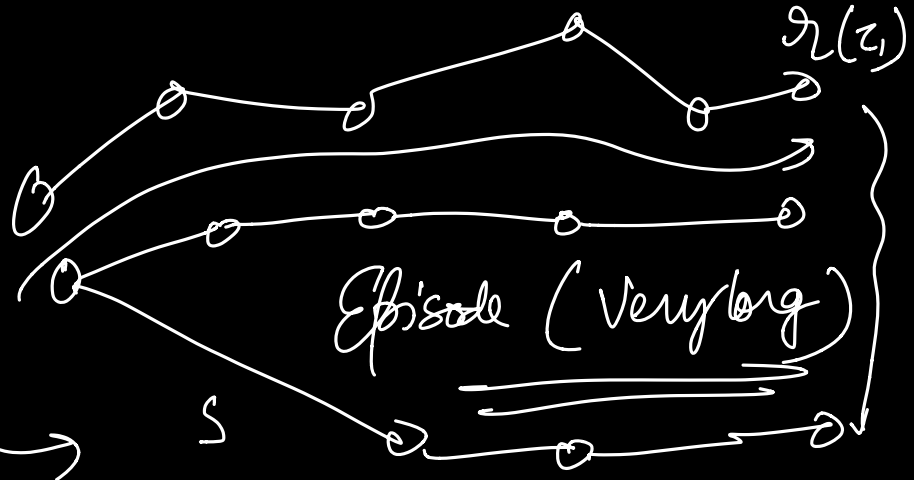
$$\nabla_\theta J(\theta) \propto \sum_{t \geq 0} r(z) \nabla_\theta \log(\pi_\theta(a_t|s_t))$$

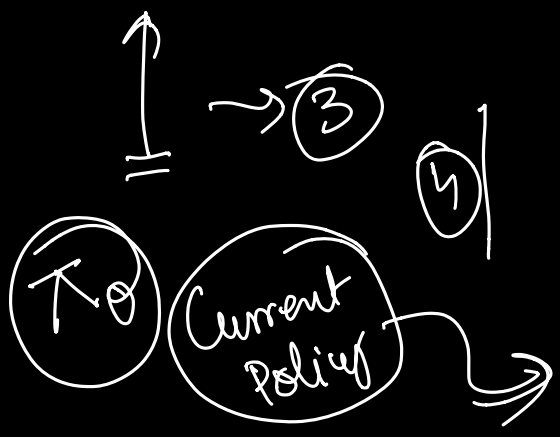
$$\sum_{t \geq 0} \left(\sum_{t' \geq t} \gamma_{t'} \right) \nabla_\theta \log(\pi_\theta(a_t|s_t))$$

$$\sum_{t \geq 0} \left(\sum_{t' \geq t} \gamma^{t'-t} \gamma_{t'} \right) * (\text{Same})$$

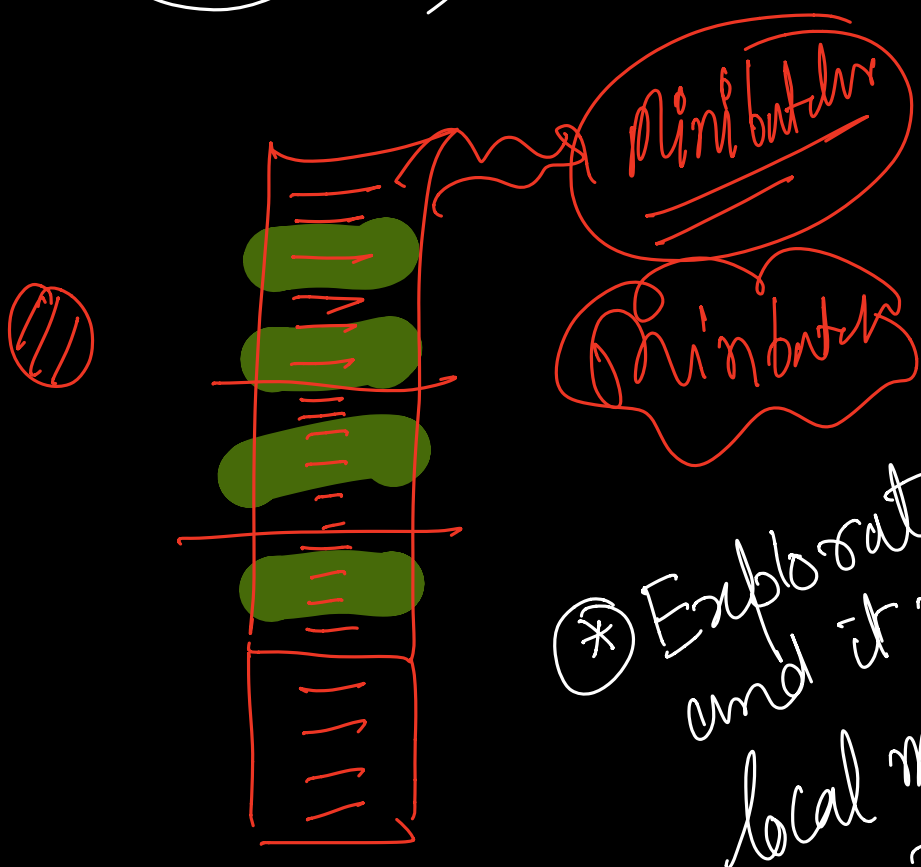
$r(z_1)$
$r(z_2)$
$r(z_3)$

Playing game





{discard}



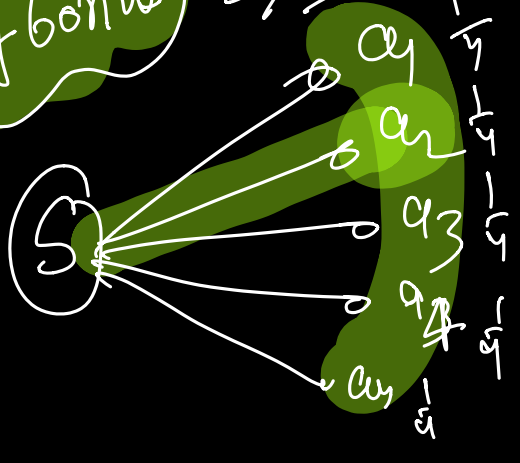
* Exploration is limited and it may stuck in local minima.

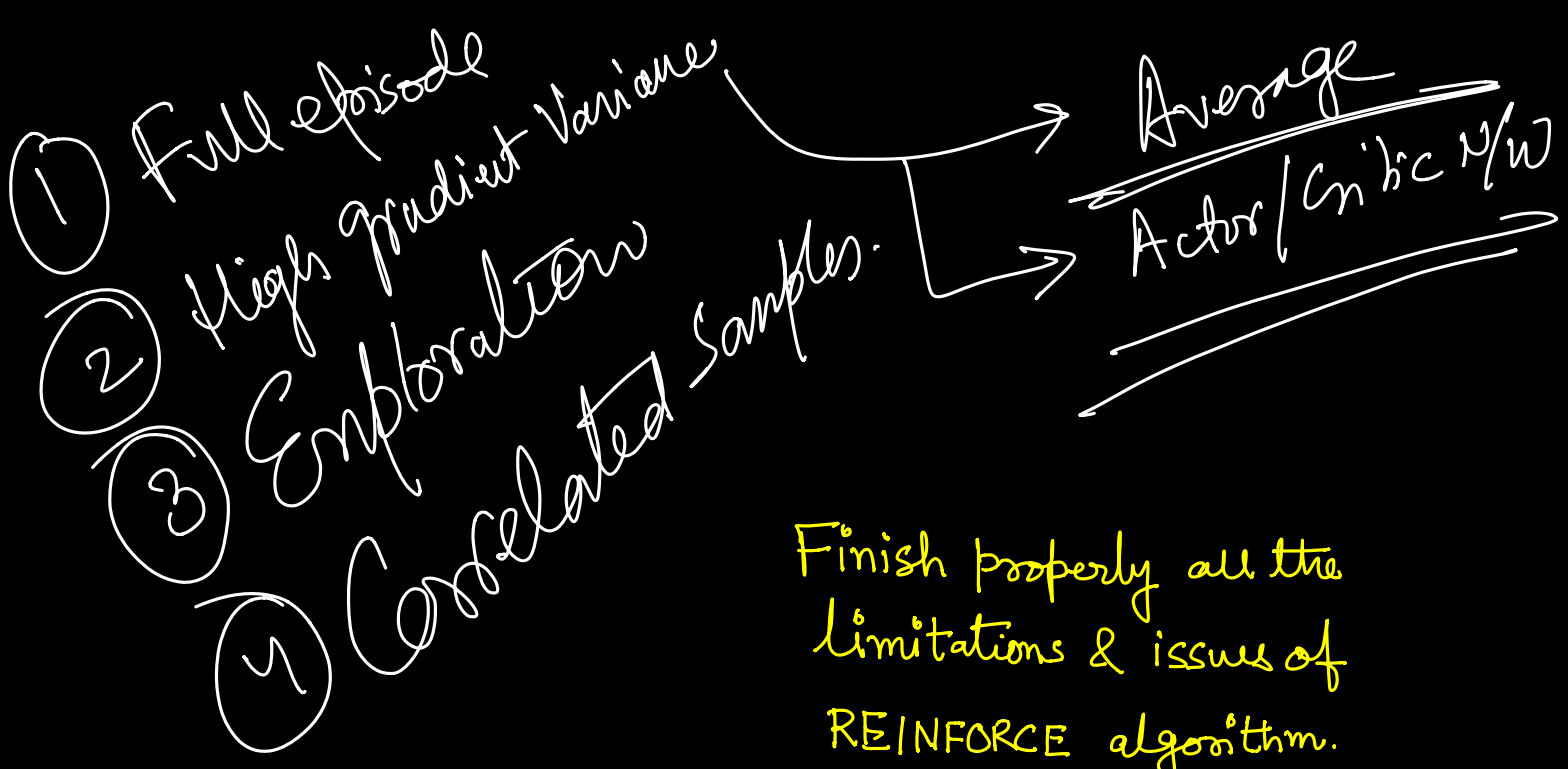
$$H(\pi) = -\sum \pi(a|s) * \log(\pi(a|s))$$

Entropy bonus

$$\Rightarrow D(QN)$$

loss fn — Entropy





Finish properly all the limitations & issues of REINFORCE algorithm.

(1) Algorithm for Policy Gradient (REINFORCE)

looping

(1) Initialize the network parameters θ.

(2) Generate (N) trajectories $\{\tau^i\}_{i=1}^{N}$ $(\tau^1, \tau^2, \tau^3 \cdots \tau^N)$ following the policy π_θ.

(3) Compute the return of the trajectory $R(\tau)$

(4) Compute the gradient:

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \right] * R(\tau)$$

(5) Update the network parameter

$$\theta = \theta + \alpha \nabla_\theta J(\theta)$$

(6) Repeat steps (2) to (5) for several iterations.

(θ) getting updated after each iterations. Hence policy got changed (become better) and then generate new data /episode/ trajectories for traing.

It is an On-policy method.

How to manage high Variance in the gradient

⊛ Actually we are using policy to generate the trajectory and then Computing $\nabla_\theta J(\theta)$ to update the policy itself.

↳ This will in turn improve the policy after each iteration.

↳ Hence returns vary greatly introducing high variance in the gradient updates.

(ii) Policy gradient with reward-to-Go

for Vanila PG

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) * R(\tau) \right]$$

Poor Credit Assignment

where $R(\tau) = \sum_{t=0}^{T-1} r_t$

but us say Reward-to-Go defined as (R_t)

Sum of the rewards of the trajectory starting from the state (S_t).

$$\therefore R_t = \sum_{t'=t}^{T-1} r(s_{t'}, a_{t'})$$

Instead of $R(z)$ use R_t

An action is only as good as the DCFR it can generated

(iii) Policy gradient with the baseline

→ Trajectory lengths are different.

→ Favoring early rewards

Can we Normalize the Reward-to-Go

↳ In order to reduce the variance one can Subtract some baseline (b) from the (R_t) (Just like DC-Component/Average)

Can we develop the notion of "Expected"

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) * (R_t - b) \right]$$

Baseline (b): It is the value that can give us the expected return from the state the agent is in. Simplest baseline can be: $b = \frac{1}{N} \sum_{i=1}^{N} R(z)$ {or even moving average}

baseline can be any function but should not dependent upon (θ). (Does not affect N/W params)

Other obvious choices can be

↳ Value function ($V(s_t)$)

Value of a state is the expected return an agent would obtain starting from the state following (π).

↳ Other options are Q fn & Advantage fn.

How can we learn the baseline functions.

↳ Just like approximating policy using (θ) network

↳ One can use value network (ϕ)

$$\nabla_\theta J(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t|s_t) \left(R_t - V(s_t)\right)^*$$

Now since the value of a state is floating point number, ϕ can be trained by minimizing M.S.E.

$(R_t) \Rightarrow$ Actual Return

$V(S_t) \Rightarrow$ Predicted Return

$$\therefore J(\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\left(R_t - V_\phi(S_t)\right)^2$$

We need to minimize $J(\phi)$ hence

$$\phi = \phi - \beta \nabla_\phi J(\phi)$$

Algorithm – REINFORCE with baseline

The algorithm of the policy gradient method with the baseline function (REINFORCE with baseline) is shown here:

1. Initialize the policy network parameter θ and value network parameter ϕ
2. Generate N number of trajectories $\{\tau^i\}_{i=1}^{N}$ following the policy π_θ
3. Compute the return (reward-to-go) R_t
4. Compute the policy gradient:

$$\nabla_\theta J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(a_t|s_t)\ (R_t - V_\phi(s_t))\right]$$

5. Update the policy network parameter θ using gradient ascent as $\theta = \theta + \alpha\nabla_\theta J(\theta)$
6. Compute the MSE of the value network:

$$J(\phi) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\left(R_t - V_\phi(s_t)\right)^2$$

7. Compute gradients $\nabla_\phi J(\phi)$ and update the value network parameter ϕ using gradient descent as $\phi = \phi - \alpha\nabla_\phi J(\phi)$
8. Repeat *steps* 2 to 7 for several iterations

(A) Advantage fn (A_t^i) :- t^{th} step of i^{th} episode.

→ How "good" it is to take an action "a", at state "s".

when $Q(s,a) \geq Value(s)$

(Action (a) is better than expected/ average.)

$$A_t^i \Rightarrow Q(s_t, a_t) - V(s_t)$$ for some policy (π).

$$\nabla_\theta J(\theta) = \sum_{t \geq 0} \left[Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t) \right] * \nabla_\theta \log \pi_\theta(a_t | s_t)$$

Scaling fn for log likelihood

(*) Minimization of the advantage function will also enforces Bellman optimality.

(*) Computation of Advantage fn needs (2) neural networks (a) for Q fn (b) for value fn [This is not optimal]

This is computationally inefficient & expensive.

Q fn, Can be

→ (a) Approximated via. Monte Carlo based Random Sampling or [Reward-as-Go].

I

$Q(S, a) \Rightarrow$ All future reward (R_t)

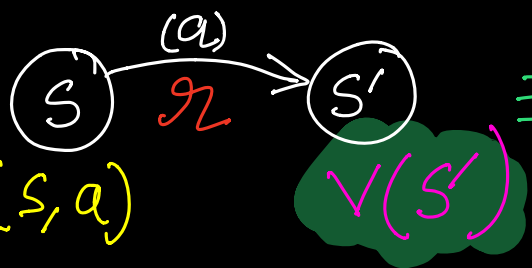
policy itself says take action (a) at (S)

(Doing it over the Experience)

(Major Concern is that, we need to have the full trajectory or trajectories)

(b) It can be approximated by using the definition of Q-fn and using Bellman optimality.

II



$$Q(S, a) = R + \gamma(V(S'))$$

This we can get from (∅)

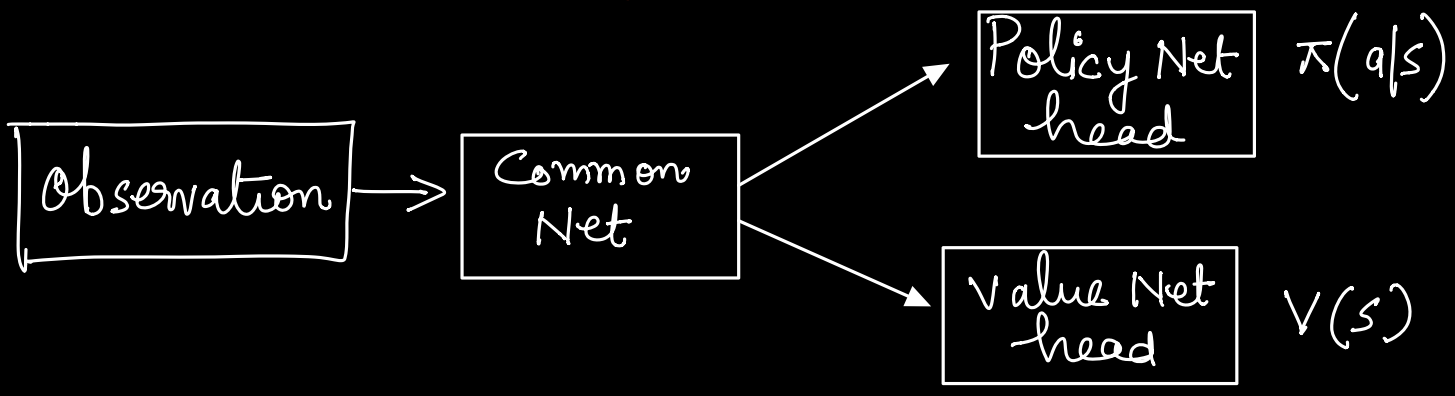
Q Value can be computed using Value fn N/W.

Now Q-fn can be computed for any step, instantly using, immediate reward & O/P of Value fn.

Hence

$$A_t = A(S, a) = R + \gamma V(S') - V(S)$$

Actor Critic Algorithm



Observation :

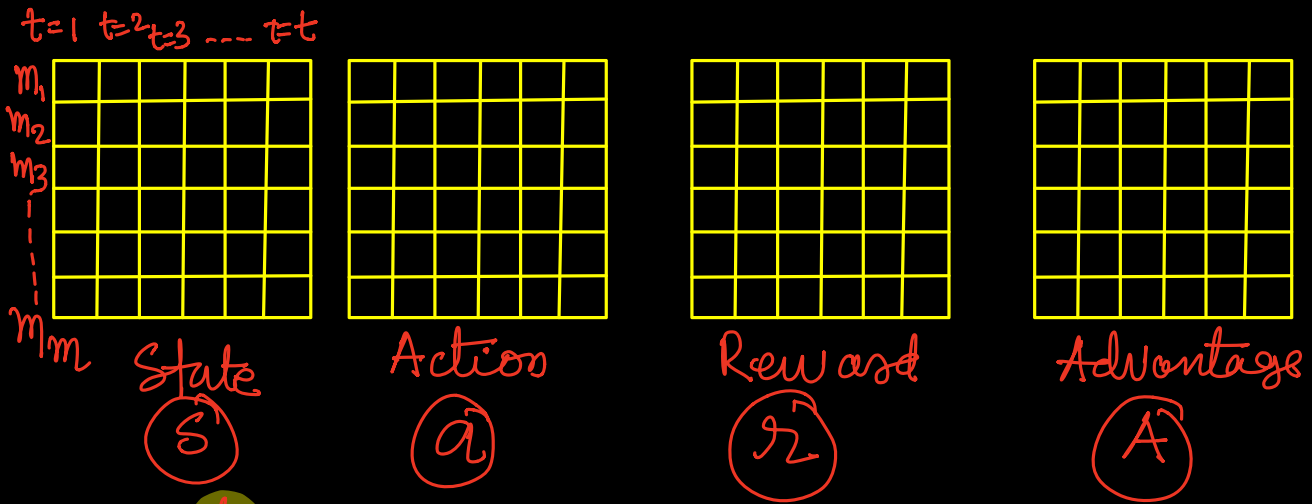
Sample (m) trajectories under the current policy (π).

m_1 $\langle s_0^1, a_0^1, r_0^1 \rangle$ $\langle s_1^1, a_1^1, r_1^1 \rangle$ ---- $\langle s_t^1, a_t^1, r_t^1 \rangle$

m_2 $\langle s_0^2, a_0^2, r_0^2 \rangle$

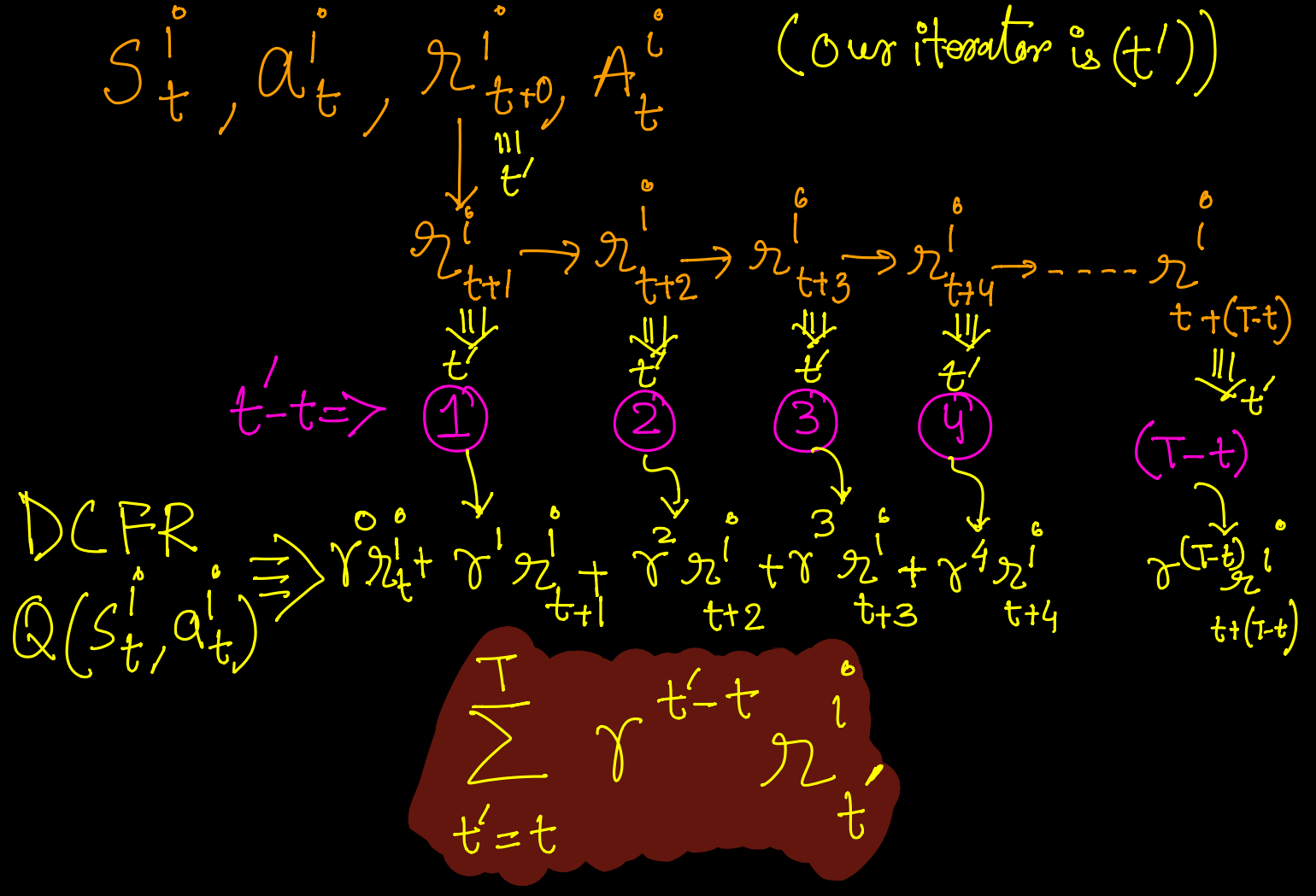
m_3 $\langle s_0^3, a_0^3, r_0^3 \rangle$

m_m $\langle s_0^m, a_0^m, r_0^m \rangle$ $\langle s_t^m, a_t^m, r_t^m \rangle$



(s) (a) (r) (A)

$S_t^i \Rightarrow$ State of the agent at t^{th} step in the (i^{th}) episode.



ALGORITHM (Actor/Critic)

(1) Initialize the
- Policy N/w $[\theta]$ (Actor)
- Q learning N/w $[\phi]$ (Critic)

(2) For each training iteration = (1, 2, --) do

After every iteration policy will be updated. New data need to be generated.

(2.1) Sample (m) trajectories under the current policy (π). (s, a, r, t)

(2.2) $\Delta\theta = 0$ — Gradient accumulator for Actor N/w. for a given policy (π), before training reset.

(2.3) for $i = 1, 2, \ldots (m)$ do — for each episode / trajectory

(2.3.1) for $t = 1, 2, \ldots (T)$ do — for each step (t)

(2.3.1.1) $$A_t^{(i)} = \sum_{t'=t}^{T} \gamma^{t'-t} * r_{t'}^{i} - V_{\phi}(s_t^i)$$

Advantage fn for the (t^{th}) step of (i^{th}) episode

It doesn't depends upon the Actor N/w (θ)

all future rewards (R_t)

Critic network

Minimization of A_t, enforces Bellman Eq.

(2.3.1.2) $$\Delta\theta = \Delta\theta + A_t^i \nabla_\theta \log(a_t^i | s_t^i)$$

Getting the updation from the i^{th} episode's t^{th} time step. scaled by (A_t^i)

Accumulating all the policy updates for (Actor N/w).

Gradient ascent

Gradient estimator step.

(2.4) $$\Delta\phi = \sum_i \sum_t \nabla_\phi \| A_t^i \|^2$$

Minimizing the advantage fn, Discounted accumulated future rewards getting close to the predicted value function.

for all episodes and for all time steps, just accumulate them all.

Critic N/w Updation

A_t minimization enforces Bellman Constraint.

(2.5) $\theta = \theta + \alpha \Delta\theta$

(2.6) $\phi = \phi - \beta \Delta\phi$

old problem of deep Q-learning:
(*) Learn Q-values for all <State, actions> pairs.

(3) End for

Also incorporate standard experience replay trick too.

Critic only learns the Q-values for generated/observed $\langle s, a \rangle$ pairs by the policy under consideration.

Are these algorithms Sample efficient (as they are on-policy)

(*) Instead of generating the full trajectory and then only compute the Return/Reward, Can we do something efficient.

(*) we can approximate (R_t) or $Q(s,a)$ as:

$$R \approx r + \gamma(V(s'))$$

(*) Now, we don't need to wait till the end of episode, to compute reward.

(*) One can compute gradient at each step & update the policy network parameters. [At each step]

Let us see the flow : π fn, V fn

1. Initialize, θ & ϕ
2. Get a random starting state (S_0)
3. Get a action $\pi(a_0|s_0)$
4. Get a reward (r_0) and next state (S_1)
5. Get the Value of (S_1) as $V_\phi(S_1)$
6. Get $Q(S_0, a_0) = r_0 + \gamma * V_\phi(S_1)$
7. Get Advantage fn, $A(S_0, a_0)$ as
$$A(S_0, a_0) = r_0 + \gamma * V_\phi(S_1) - V_\phi(S_0)$$
8. Get Policy Gradient : $\nabla_\theta J(\theta)$
$$\nabla_\theta J(\theta) = \nabla_\theta \log \pi_\theta(a_0|s_0) \left(r_0 + \gamma * V_\phi(S_1) - V_\phi(S_0)\right)$$
9. Update Actor N/W (θ) : $\theta = \theta + \alpha \nabla_\theta J(\theta)$
(See NO $(\Delta\theta)$ accumulation) (Gradient Ascent)
10. Get Critic N/W loss : $J(\phi) = A(S_0, a_0) = r_0 + \gamma * V_\phi(S_1) - V_\phi(S_0)$
(ϕ)
11. Get $\nabla_\phi J(\phi)$, Value Gradient and update Critic N/W (using gradient descent) : $\phi = \phi - \alpha \nabla_\phi J(\phi)$

The actor-critic algorithm

The steps for the actor-critic algorithm are:

1. Initialize the actor network parameter θ and the critic network parameter ϕ
2. For N number of episodes, repeat *step 3*
3. For each step in the episode, that is, for $t = 0, \ldots, T\text{-}1$:
    1. Select an action using the policy, $a_t \sim \pi_\theta(s_t)$
    2. Take the action a_t in the state s_t, observe the reward r, and move to the next state s'_t
    3. Compute the policy gradients:

$$\nabla_\theta J(\theta) = \nabla_\theta \log \pi_\theta (a_t|s_t)\big(r + \gamma V_\phi(s'_t) - V_\phi(s_t)\big)$$

    4. Update the actor network parameter θ using gradient ascent:

$$\theta = \theta + \alpha \nabla_\theta J(\theta)$$

    5. Compute the loss of the critic network:

$$J(\phi) = r + \gamma V_\phi(s'_t) - V_\phi(s_t)$$

    6. Compute gradients $\nabla_\phi J(\phi)$ and update the critic network parameter ϕ using gradient descent:

$$\phi = \phi - \alpha \nabla_\phi J(\phi)$$

Both updates are done at each & every step of the episode

Sample efficient.