AMRITA VISHWA VIDYAPEETHAM

21AIE311 - REINFORCEMENT LEARNING

ELEMENTS OF RL

A A NIPPUN KUMAAR
DEPARTMENT OF CSE
AMRITA SCHOOL OF ENGINEERING, BANGALORE

PREVIOUSLY

- Course Overview
- Reinforcement Learning Defined
- Types of Machine Learning
- Why Reinforcement Learning
- Interesting Applications

LECTURE OVERVIEW

- RL - What ?? Why ??
- Key Features of RL
- History - Timeline
- Real World RL Examples
- Elements of RL
- An Example

WHAT IS RL ?

- Learning from interaction
- Goal-oriented learning
- Learning about, from, and while interacting with an external environment
- Learning what to do—how to map situations to actions—so as to maximize a numerical reward signal

WHY RL?

- Learn to make good sequences of decisions
  - Repeated Interactions with World – Sequences of decisions
  - Reward for sequence of Decisions – Good
- How the world works in advance is unknown? – Learn

KEY FEATURES OF RL

- Learner is not told which actions to take
- Trial-and-Error search
- Possibility of delayed reward (sacrifice short-term gains for greater long-term gains)
- The need to explore and exploit
- Considers the whole problem of a goal-directed agent interacting with an uncertain environment

HISTORY – TIMELINE

- DeepMind 2015 – Mastered 49 Atari Games – Won human experts in 29 games – Deep Q network- used pixels and game score as inputs
- Deep Mind – DQN
  - Google DeepMind 2016 – AlphaGo – 99.8% winning Rate – won human champion by 5-0
    - Value networks to evaluate board positions and Policy networks to select moves
  - Google DeepMind 2017 – AI Managed to learn how to walk, run and climb without any prior guidance
    - https://www.youtube.com/watch?v=gn4nRCC9TwQ
- OpenAI – Dactyl is a system for manipulating objects using a Shadow Dexterous Hand – Trained entirely in simulation and transfers its knowledge to reality – First RL to be working in the real world
- NeuroIPS 2019 – Minecraft Reinforcement Learning Competition
- Multiagent
  - https://www.youtube.com/watch?v=kopoLzvh5jY

REAL WORLD RL EXAMPLES

- A Master Class Chess Player
  - The choice is informed both by planning– anticipating possible replies and counter-replies.
  - By immediate, intuitive judgments of the desirability of particular positions and moves.
- A gazelle calf
  - Struggles to its feet minutes after being born.
  - Half an hour later it is running at 20 miles per hour.

REAL WORLD RL EXAMPLES

- Phil prepares his breakfast.
- A complex web of conditional behavior and interlocking goal–subgoal relationships : walking to the cupboard, opening it, selecting a cereal box, then reaching for, grasping, and retrieving the box.
- Other complex, tuned, interactive sequences of behavior are required to obtain a bowl, spoon, and milk carton.
- Each step involves a series of eye movements to obtain information and to guide reaching and locomotion.
- Rapid judgments are continually made about how to carry the objects or whether it is better to ferry some of them to the dining table before obtaining others.
- Each step is guided by goals, such as grasping a spoon or getting to the refrigerator, and is in service of other goals, such as having the spoon to eat with once the cereal is prepared and ultimately obtaining nourishment.
- Whether he is aware of it or not, Phil is accessing information about the state of his body that determines his nutritional needs, level of hunger, and food preferences.

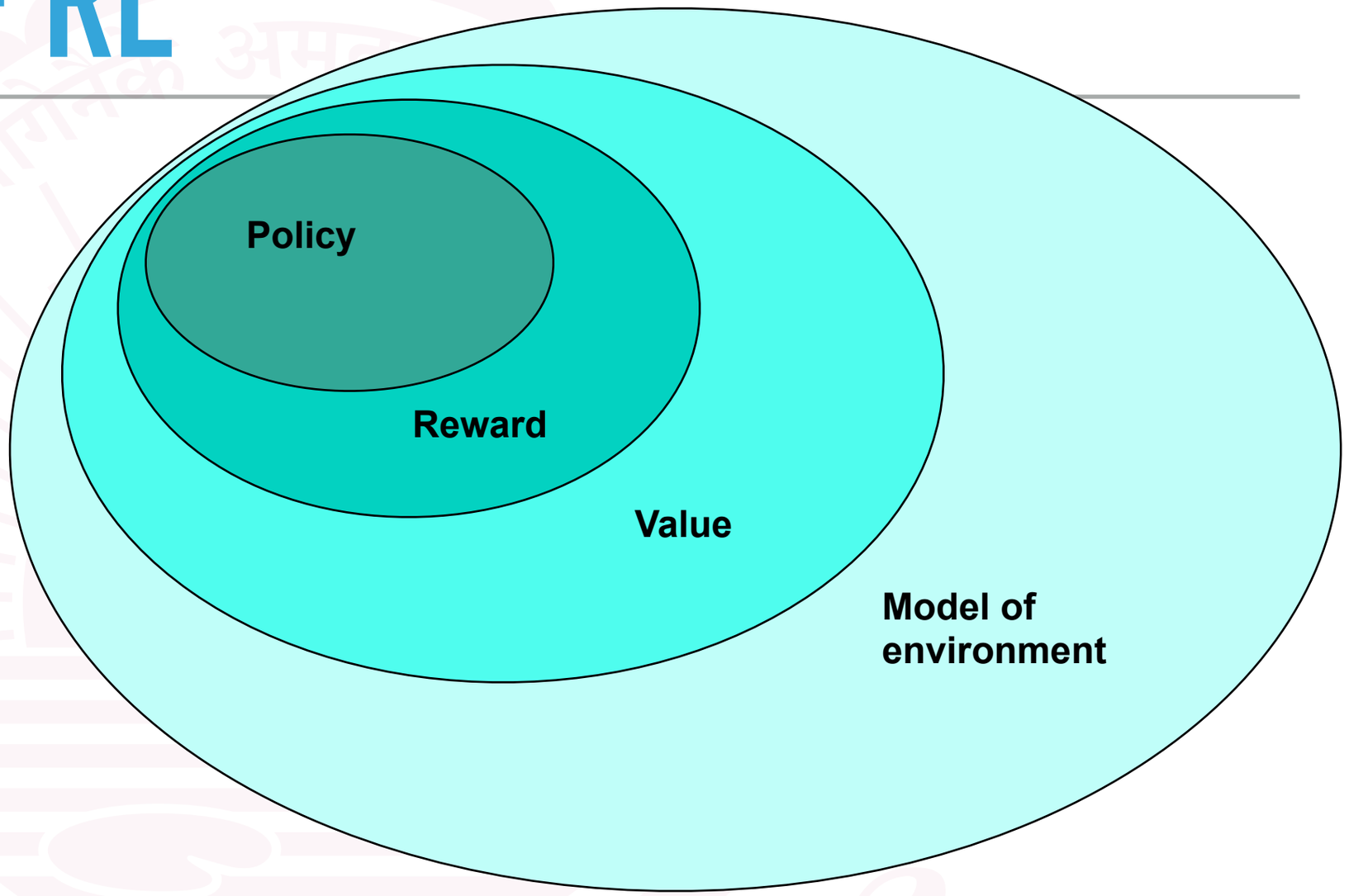
REAL WORLD RL EXAMPLES

- In all these examples
  - Involves Interaction between active decision making agent and its environment.
  - Within which the agent seeks to achieve a goal despite uncertainty in the environment.
  - The agents actions are permitted to affect the future state of the environment, thereby affecting the actions and opportunities available to the agent at later times.
- Correct choice requires taking into account indirect, delayed consequences of actions, and thus may require foresight or planning.
- The effects of actions cannot be fully predicted; thus the agent must monitor its environment frequently and react appropriately.
  - For example, Phil must watch the milk he pours into his cereal bowl to keep it from overflowing.

REAL WORLD RL EXAMPLES

- In all these examples
  - Involve goals that are explicit in the sense that the agent can judge progress toward its goal based on what it can sense directly.
  - The agent can use its experience to improve its performance over time.
- The knowledge the agent brings to the task at the start—either from previous experience with related tasks or built into it by design or evolution—influences what is useful or easy to learn, but interaction with the environment is essential for adjusting behavior to exploit specific features of the task.

ELEMENTS OF RL



- Policy: what to do ?
- Reward: what is good ?
- Value: what is good because it predicts reward ?
- Model: what follows what ? (Optional)

ELEMENTS OF RL

- Policy
  - A policy defines the learning agent's way of behaving at a given time.
  - Roughly speaking, a policy is a mapping from perceived states of the environment to actions to be taken when in those states.
  - It corresponds to what in psychology would be called a set of stimulus–response rules or associations.
- In some cases the policy may be a simple function or lookup table, whereas in others it may involve extensive computation such as a search process.
- The policy is the core of a reinforcement learning agent in the sense that it alone is sufficient to determine behavior.
- In general, policies may be stochastic, specifying probabilities for each action.

ELEMENTS OF RL

- Reward
  - A reward signal defines the goal of a reinforcement learning problem.
  - On each time step, the environment sends to the reinforcement learning agent a single number called the reward.
  - The agent's sole objective is to maximize the total reward it receives over the long run.
  - The reward signal thus defines what are the good and bad events for the agent.
  - In a biological system, we might think of rewards as analogous to the experiences of pleasure or pain.
  - In general, reward signals may be stochastic functions of the state of the environment and the actions taken.

ELEMENTS OF RL

- Value
  - Whereas the reward signal indicates what is good in an immediate sense, a value function specifies what is good in the long run.
  - Roughly speaking, the value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state.
  - For example, a state might always yield a low immediate reward but still have a high value because it is regularly followed by other states that yield high rewards. Or the reverse could be true.
  - Action choices are made based on value judgments. We seek actions that bring about states of highest value, not highest reward, because these actions obtain the greatest amount of reward for us over the long run.
  - In fact, the most important component of almost all reinforcement learning algorithms we consider is a method for efficiently estimating values.

ELEMENTS OF RL

- Model
  - This is something that mimics the behavior of the environment, or more generally, that allows inferences to be made about how the environment will behave.
  - For example, given a state and action, the model might predict the resultant next state and next reward.
- Methods for solving reinforcement learning problems that use models and planning are called model-based methods, as opposed to simpler model-free methods that are explicitly trial-and-error learners - viewed as almost the opposite of planning.
- Modern reinforcement learning spans the spectrum from low-level, trial-and-error learning to high-level, deliberative planning.

EXAMPLE – TIC-TAC-TOE

- **Game Rules**
- Two players take turns playing on a three-by-three board.
- One player plays Xs and the other Os until one player wins by placing three marks in a row, horizontally, vertically, or diagonally.
- If the board fills up with neither player getting three in a row, then the game is a draw.

X	O	O
O	X	X
		X

https://playtictactoe.org

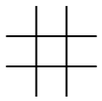
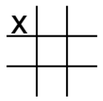
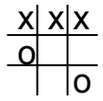
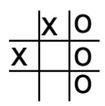
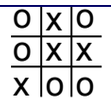
EXAMPLE – TIC-TAC-TOE

- A skilled player can play so as never to lose, let us assume that we are playing against an imperfect player, one whose play is sometimes incorrect and allows us to win.
- For the moment, in fact, let us consider draws and losses to be equally bad for us.
- How might we construct a player that will find the imperfections in its opponent's play and learn to maximize its chances of winning?

EXAMPLE – TIC-TAC-TOE

- Steps to deal with this problem using value function based RL.
- Step-1 - Initializing value table
  - Setup a table with each possible states of the game and its corresponding value function
    - 3^(3*3) = 19683
- Each number will be the latest estimate of the probability of our winning from that state and the whole table is the learned value function.

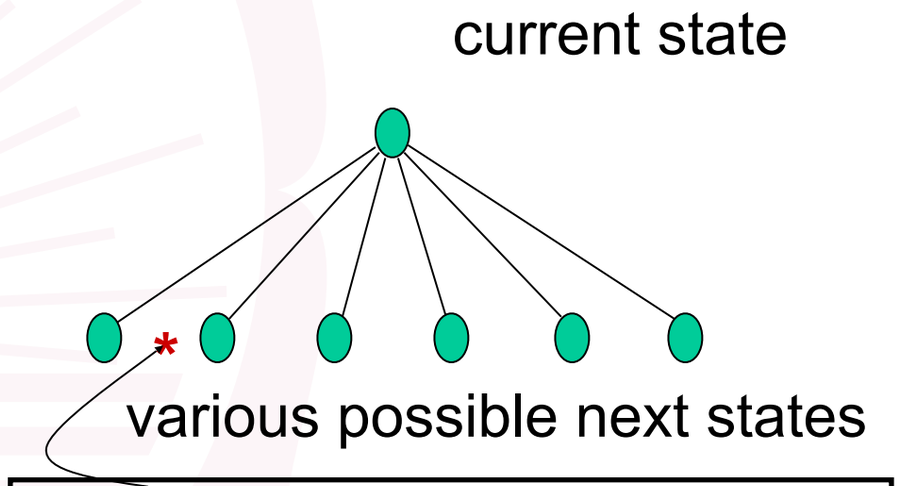
EXAMPLE – TIC-TAC-TOE

State	$V(s)$ – estimated probability of winning	
	.5	?
	.5	?
⋮	⋮	
	1	win
⋮	⋮	
	0	loss
⋮	⋮	
	0	draw

- Initial States Probability
  - Win states - 1
  - Draw and other states - 0.5
  - Lose - 0
- As the agent plays the probability values will be changed based on experience

EXAMPLE – TIC-TAC-TOE

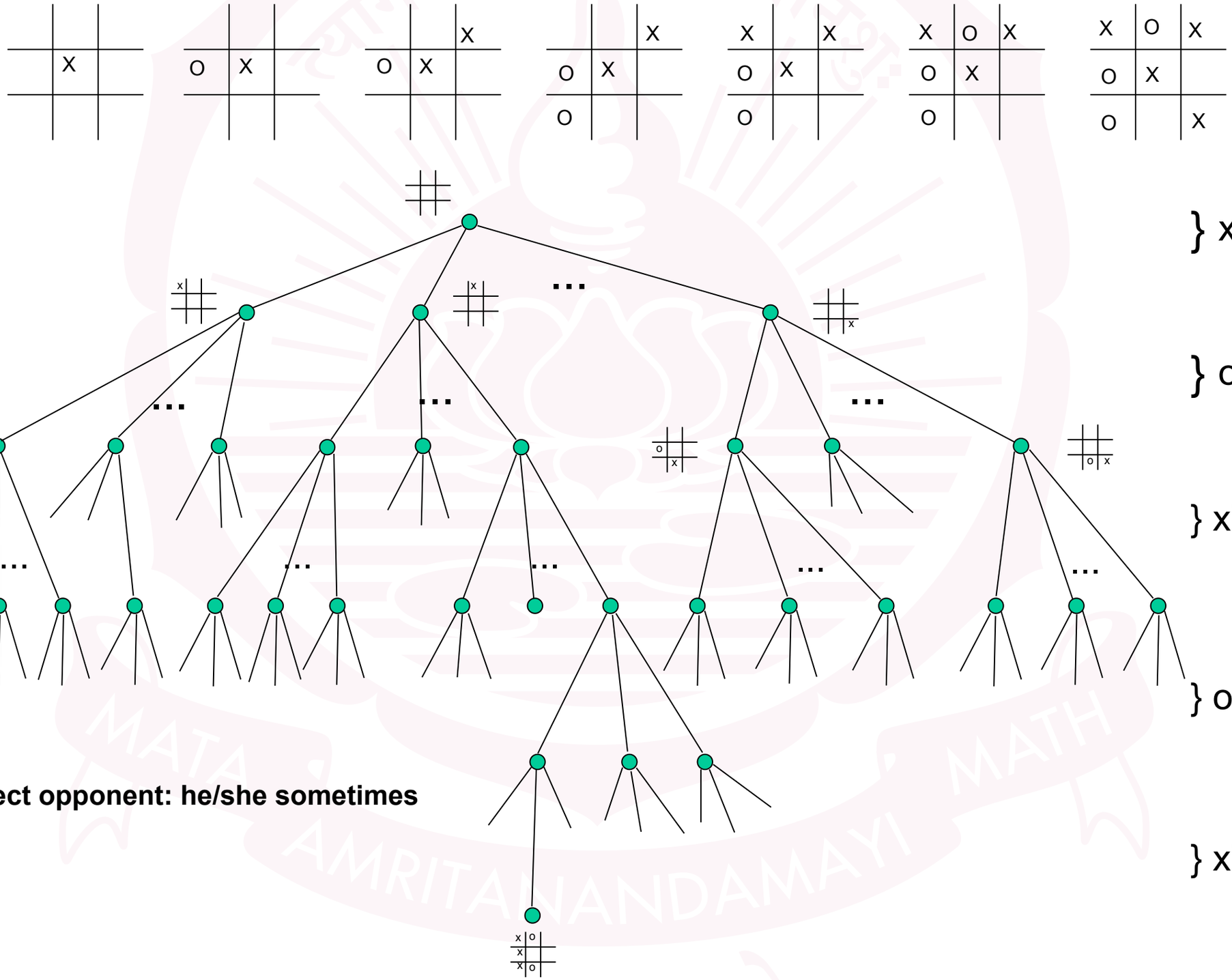
- Step-2 - Start playing
  - We then play many games against the opponent.
  - To select the next move - examine all the possible next states (one for each black space on the board) and its value in the table.
- Exploration vs Exploitation
  - Exploration - Make a random move irrespective of the recent value.
  - Exploitation - Choose a move based on the hight value among the next possible states.



Just pick the next state with the highest estimated prob. of winning — the largest $V(s)$; a ***greedy*** move.

But 10% of the time pick a move at random; an ***exploratory move***.

EXAMPLE – TIC-TAC-TOE



Assume an imperfect opponent: he/she sometimes makes mistakes

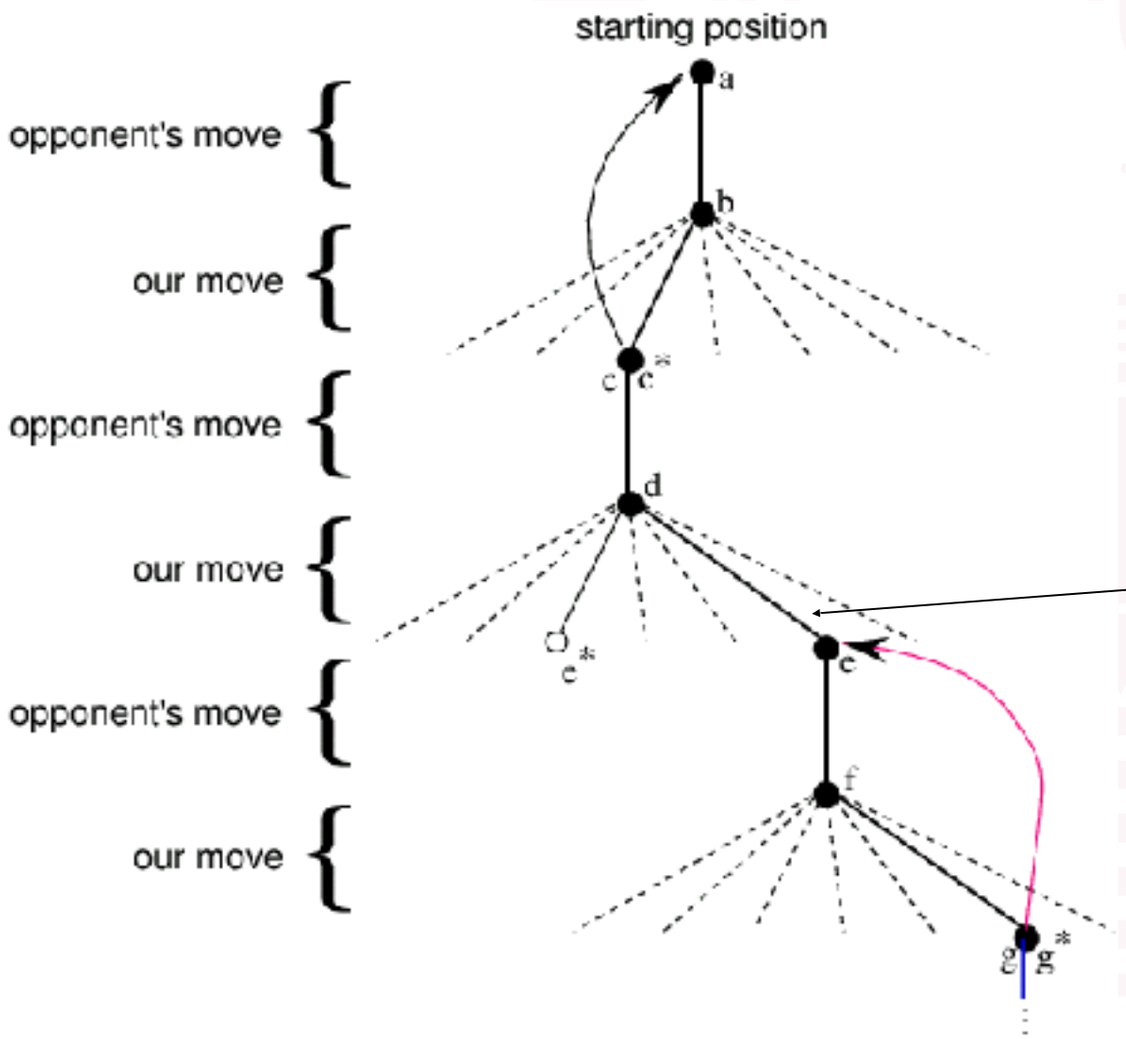
EXAMPLE – TIC-TAC-TOE

- Step-3 - Updating the value table (Learning)
  - While we are playing, we change the values of the states in which we find ourselves during the game. We attempt to make them more accurate estimates of the probabilities of winning.
  - To do this, we "back up" the value of the state after each greedy move to the state before the move.
- The current value of the earlier state is updated to be closer to the value of the later state.
- This can be done by moving the earlier state's value a fraction of the way toward the value of the later state.

$$V(S_t) \leftarrow V(S_t) + \alpha\Big[V(S_{t+1}) - V(S_t)\Big]$$

TEMPORAL-DIFFERENCE LEARNING

EXAMPLE – TIC-TAC-TOE



s – *the state before our greedy move*

s' – *the state after our greedy move*

We increment each $V(s)$ *toward* $V(s')$ – *a* ***backup*** *:*

$$V(s) \leftarrow V(s) + \alpha\left[V(s') - V(s)\right]$$

α *a small positive fraction, e.g.,* $\alpha = .1$

the ***step - size parameter***

EXAMPLE – TIC-TAC-TOE

- Outcome
  - If the step-size parameter is reduced properly over time, then this method converges, for any fixed opponent.
  - If the step-size parameter is not reduced all the way to zero over time, then this player also plays well against opponents that slowly change their way of playing.

EXAMPLE – TIC-TAC-TOE

- This example illustrates all the key features of a RL discusses earlier
  - Emphasis on learning while interacting with an environment
  - There is a clear goal, and correct behavior requires planning or foresight
  - The simple reinforcement learning player would learn to set up multi-move traps for a shortsighted opponent.
  - It is a striking feature of the reinforcement learning solution that it can achieve the effects of planning and lookahead without using a model of the opponent and without conducting an explicit search over possible sequences of future states and actions.

OTHER IMPORTANT FEATURES

- Tic-tac-toe is a small, finite state set,
  - whereas reinforcement learning can be used for state set which is very large or even infinite.
  - Backgammon – No of states 1020
- Tic-tac-toe – No prior information ,
  - however in RL prior information can be incorporated
- Tic-tac-toe – Access to true state,
  - however in RL is applied to part of the state which is hidden or when different states appear to the learner to be same

THANK YOU !!!