FINITE MARKOV DECISION PROCESSES

- STATE, REWARD, ACTION • AGENT/ENVIRONMENT • **POLICY** π_T, $\pi_T(a|s)$ PROBABILITY, MAPS STATES TO PROBABILITY OF SELECTING ACTION
- EXPRESS THE GOAL THROUGH **REWARD SIGNAL**: WHAT TO ACHIEVE NOT HOW TO ACHIEVE
- **CUMULATIVE REWARD**: A SPECIFIC FUNCTION OF THE REWARD SEQUENCE, IE SUM
- **EPISODIC TASKS**: THERE IS A TERMINATING STATE • **CONTINUING TASKS**: NO TERMINAL STATE → **DISCOUNTED RETURNS**: $G_T = R_{T+1} + \gamma R_{T+2} + \gamma^2 R_{T+3} + \sum \gamma^k R_{T+k+1}$
  ↳ CAN BE SEEN AS CONTINUING WHERE TERMINAL IS A ABSORBING STATE WITH $R=0$

  $0 < \gamma < 1$ DISCOUNT RATE
  - 0 ONLY IMMEDIATE REWARDS
  - 1 FARSIGHTEDNESS
- **GENERAL RETURN** $G_T = \sum \gamma^k R_{T+k+1}$ • **STATE SIGNAL**: USUALLY SET TO HAVE MARKOV PROPERTY
- RL TASK + MARKOV PROPERTY → FINITE MARKOV DECISION PROCESS
- DIAGRAM WITH STATE + ACTION NODES. $\sum P_{OUT}(a) = 1$

  EXPECTED REWARDS $R(s,a) = \sum R \sum_{s' \in S} p(s', R|s,a)$

  STATE TRANSITIONS: $p(s'|s,a) = \sum p(s', R|s,a)$

  EXPECTED REWARDS S-A-NS: $R(s,a,s') = \dfrac{\sum_R R \cdot p(s', R|s,a)}{p(s'|s,a)}$

- **VALUE FCN**: $V_\pi(s) = E_\pi[G_T | S_T = s] = E_\pi[\sum \gamma^k R_{T+k+1} | S_T = s]$ → VALUE OF STATE UNDER POLICY π, EXPECTED RETURN OF STARTING IN S AND FOLLOWING π

  STATE-VALUE FCN FOR POLICY π

- $Q_\pi(s,a) = E_\pi[G_T | S_T = s, A_T = a] = E_\pi\left[\sum \gamma^k R_{T+k+1} \middle| S_T = s, A_T = a\right]$ → VALUE OF TAKING A, UNDER STATE S, AND FOLLOWING π

  ACTION-VALUE FCN FOR POLICY π

BELLMAN EQUATION: $V_\pi(s) = \ldots = \sum_a \pi(a|s) \sum_{s',R} p(s', R|s,a)[R + \gamma V_\pi(s')]$ RELATIONSHIP BETWEEN VALUE OF STATES AND ITS VALUES OF SUCCESSORS

• FUNDAMENTAL IDENTITY $\approx T^\pi V^\pi = V^\pi$, T^π BELLMAN OPERATOR
↳ CONTRACTION FOR LEAST POINTS OF γ → CONVERGENCE UNIQUE SOLUTION

OPTIMAL VALUE FCN: FINITE MDP HAVE CLOSED-FORM OPTIMAL POLICY. VALUE FCN INDUCE PARTIAL ORDERING OVER POLICIES.

$\pi \geq \pi'$ IFF $V_\pi(s) \geq V_{\pi'}(s)$ $\forall s \in S$. MULTIPLE OPTIMAL POLICIES ⟷ EQUIVALENT, SAME $V_*(s)$ AND $Q_*(s)$

$V_*(s) = \max_\pi V_\pi(s)$ $\quad Q_*(s,a) = \max_\pi Q_\pi(s,a) = E[R_{T+1} + \gamma V_*(S_{T+1}) | S_T = s, A_T = a]$

BELLMAN OPTIMALITY EQUATION:

$$\begin{cases} V_*(s) = \max \sum p(s', R|s,a)[R + \gamma V_*(s')] \\ Q_*(s,a) = \sum_{s,R} p(s', R|s,a)[R + \gamma \max Q_*(s', a')] \end{cases}$$

- FOR FINITE MDP, UNIQUE SOLUTION INDEPENDENT OF POLICY
- VALUE OF STATE UNDER OPTIMAL POLICY MUST EQUAL EXPECTED RETURN FOR BEST ACTION FROM THAT STATE
- IS $|S|$ SYSTEM OF EQUATION IN $|S|$ UNKNOWNS, CAN SOLVE IF DYNAMICS ARE KNOWN • **FROM V_*** BEST ACTIONS AFTER 1-STEP SEARCH, OPTIMAL IN LONG RUN → GREEDY BUT OPTIMAL
- **FROM Q_*** EVEN EASIER, FOR ANY S → IS A ARGMAX $Q_*(s,a)$ NOT EVEN 1 STEP SEARCH.
  IS LIKE ALREADY CACHING THE RESULTS. WE DON'T HAVE TO KNOW ANYTHING ABOUT FUTURE STATES AND THEIR VALUE, NO DYNAMICS NEEDED!!!
- **IN PRACTICE**: APPROXIMATIONS BECAUSE OPTIMAL COMPUTATIONS REQUIRE EXHAUSTIVE SEARCH. BOOM. RL IS ONLINE → WE CAN 'PRUNE' LOW OCCURRING STATES.
  ↳ AND COMPLETE KNOWLEDGE