PLANNING AND LEARNING

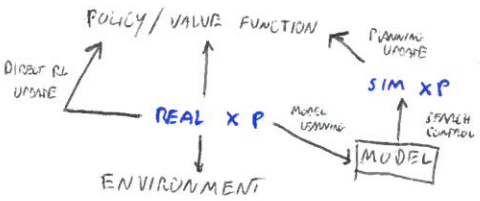
MODEL: SIMULATED EXPERIENCE, SAMPLE MODEL OR DISTRIBUTION MODEL

PLANNING: MODEL $\xrightarrow{\text{PLANNING}}$ NEW/IMPROVED POLICY. STATE-SPACE VS PLAN-SPACE PLANNING.

GENERAL STRUCTURE: MODEL $\longrightarrow$ SIMULATED EXPERIENCE $\xrightarrow{\text{BACKUPS}}$ VALUES $\longrightarrow$ POLICY

- WHO CARES IF EXPERIENCE IS REAL OR SIMULATED $\rightarrow$ PLANNING $\leftrightarrows$ LEARNING.

DYNA-Q (DYNA AGENTS) • USUALLY SAME RL METHOD • PLANNING/SIM ONLY VISITS PREVIOUSLY OBSERVED STATES AND PREVIOUSLY TAKEN ACTIONS IN S. RANDOM OFC.



- DYNA Q+: Q + EXPLORATION BONUS
- DYNA-AC: ACTOR-CRITIC

WHEN ENVIRONMENT IS STOCHASTIC / CHANGES OVER TIME:

- **OPTIMISTIC MODEL:** CHANGES EASILY DISCOURAGED BECAUSE EXPLOITABLES RESULT NOW IN WORSE PERFORMANCE.
- **PESSIMISTIC MODEL:** VERY DIFFICULT TO FIND NEW OPPORTUNITIES. • KEEP TRACK OF HOW LONG OF S-A PAIRS HAVE BEEN ATTEMPTED → GIVE BONUS POINTS FOR SIMMING THEM

PRIORITIZED SWEEPING:

- UNIFORM SELECTION OF S-A PAIRS TO SIM IS NOT EFFICIENT: WASTEFUL/WASTED BACKUPS.
- **IDEA:** LET'S WORK BACKWARDS FROM STATES WHOSE VALUE HAS CHANGED (IN REAL ENVIRONMENT) AND PROPAGATE TO PREDECESSORS. BACKUP ACTIONS LEADING INTO THAT... KEEP TRACK OF MAGNITUDE OF CHANGES. PRIORITIZE BACKUPS ABT THAT. THRESHOLD.
- VALID IN PRINCIPLE BUT VERY HARD TO IMPLEMENT IN PRACTICE, ALSO RESTRICTED TO FEW CASES

BACKUPS:

- {STATE, VALUES, ACTION VALUES} {OPTIMAL POLICY, ARBITRARY POLICY} {FULL, SAMPLE}
- 7 OF 8 COMBOS ARE USEFUL ALGOS
- BRANCHING FACTOR: NO. OF POSSIBLE NEXT STATES FOR (S, A)
- $b = 1$ SAMPLE VS FULL BACKUPS ARE SAME
- FOR MODERATELY LARGE b, SAMPLE BACKUPS OFFER MORE (s/A) IMPROVEMENT PER UNIT OF COMPUTATION TIME. 'AVALANCHE' EFFECT OF ERROR ESTIMATES TO SUCCESSOR STATES

TRAJECTORY SAMPLING

HOW TO DISTRIBUTE BACKUPS)

EXHAUSTIVE SWEEP (EVERY STATE): VERY EXPENSIVE, UNFOCUSED, INEFFICIENT

UNIFORM SWEEP: EFFICIENT BUT STILL UNFOCUSED

ACCORDING TO ON-POLICY DISTRIBUTION: FOLLOWING THE CURRENT POLICY. INTERACT WITH MODEL AND FOLLOW CURRENT π, KEEP ON SAMPLING ACTIONS. → **TRAJECTORY SAMPLING**

- EVEN IF WE COULD COMPUTE EXPLICITLY THE DISTRIBUTION, IT WOULD BE INEFFICIENT

\+ HELPS SHORT-TERM: ESPECIALLY IF MANY STATES AND LOW b

− MAY HURT LONG-TERM BECAUSE MORE LIKELY STATES ALREADY HAVE CORRECT VALUES.

HEURISTIC SEARCH

- EXTENDS GREEDY SEARCH BEYOND SINGLE STEP
- IF UNTIL END OF EPISODE → WE RETRIEVE OPTIMAL ACTION
- STATE-SPACE SEARCH AS MULTIPLE ONE-STEP BACKUP PIECED TOGETHER.
- DEPTH/SPEED TRADEOFF
- CAN ALSO USE TO DISTRIBUTE BACKUPS → USE LIKELY HOPS OF ACTIONS TO ALLOCATE BACKUPS. (IE BY COMPARING WITH TOTAL STATE VALUE)
  → MEMORY AND CPU TIME FOCUS
- GO ON STATES, ACTIONS IMMEDIATELY DEEPER FROM CURRENT SPACE