EXTRA ARTICLES

DM-09: DETERMINISTIC POLICY GRADIENTS ALGORITHMS

- THEY EXIST, NOT FOUND BEFORE. FOLLOW GRADIENT OF ACTION-VALUE FCN + INTEGRATES OVER STATE SPACE ONLY
- ON-POLICY / OFF-POLICY DETERMINISTIC ACTOR-CRITIC. • ANALOGOUS TO Q-LEARNING IN POLICY GRADIENT CONTEXTS
- $a = \mu_\theta(s)$

DM-15: BAYES-ADAPTIVE SIMULATION-BASED SEARCH WITH VALUE FUNCTION APPROXIMATION.

- FCN APPROX TO ESTIMATE VALUE OF INTERACTION HISTORIES DURING SEARCH → GENERALIZATION. FCN APPROX FOR VALUES OF POSSIBLE SIMULATED HISTORIES.
- IMPORTANCE SAMPLING OF POSTERIOR SAMPLES TO COMPRESS HISTORIES INTO FINITE-DIM VECTOR
- UNCERTAINTY ON MODEL DYNAMICS → CERTAINTY ABOUT CURRENT STATE IN AUGMENTED SPACE (POSSIBLE HISTORIES)
- SIMULATIONS ARE RUN FROM CUR BELIEF STATE → ONLINE PARAM UPDATES ON SIM EXPERIENCE
- HISTORIES WITH SAME/SIMILAR BELIEF → SAME/SIMILAR REPRESENTATIONS

DM-23 COMPRESS AND CONTROL

POLICY EVALUATION BY LEARNING TIME-INDEPENDENT STATE-ACTION CONDITIONAL DISTRIBUTIONS. THESE ARE COMPRESSED.
$P(z|s,A)$

DM-26 TOWARDS MINIMAX OFF-POLICY VALUE ESTIMATION

MATH. HARD. IMPORTANCE/WEIGHTED IMPORTANCE SAMPLING AREN'T SO GOOD

DM-41 VARIATIONAL INFORMATION MAXIMIZATION FOR INTRINSICALLY MOTIVATED RL

MUTUAL INFORMATION → EMPOWERMENT. INTRINSIC MOTIVATION. CHANNEL CAPACITY OF INFO IN ACTION SEQUENCE ABOUT FUTURE STATE.
FOR INTERNAL PLANNING, EXPLORATION POLICY

EXTERNAL + INTERNAL (CRITIC) ENVIRONMENT. EVERYTHING IS A DEEP NETWORK. USE STANDARD ALGO FOR BEHAVIOR POLICIES.
VISION, MAZE TASKS

DM-44 USING LOCALIZATION AND FACTORIZATION TO REDUCE THE COMPLEXITY OF REINFORCEMENT LEARNING

AIXI STUFF. HUTTER & SUNEHAG. NO IDEA

DM-51 RATIONALITY, OPTIMISM AND GUARANTEES IN GENERAL RL

AIXI HUTTER & SUNEHAG. JOURNAL VERSION. 50 PP MONSTER.
FRAMEWORK FOR GENERAL RL WITH RATIONALITY AXIOMS FOR OPTIMISM → CRUCIAL FOR SYSTEMATIC EXPLORATIVE BEHAVIOR.

DM-53 LEARNING CONTINUOUS CONTROL POLICIES BY STOCHASTIC VALUE GRADIENTS.

ESTIMATES (VIA A DIFFERENTIABLE ENVIRONMENT MODEL) POLICY, MODEL, REWARD FCN VIA BACKPROP.
ANALYTIC POLICY GRADIENT BY BACKPROP OF REWARD ALONG TRAJECTORY → VALUE GRADIENT
STOCHASTIC VALUE GRADIENT BPROP THROUGH STOCHASTIC BELLMAN EQUATION → REPARAM TRICK
NNETS AS MODEL. JOINT MODEL + POLICY TRAINING. SVG ∞, SVG 1, SVG 0 ALGOS. — MODEL FREE. ANALOGUE OF DETERMINISTIC POLICY GRADIENT
ON POLICY FINITE HORIZON, OFF POLICY W REPLAY

MOAR-11 CHANGING THE ENVIRONMENT BASED ON EMPOWERMENT AS INTRINSIC MOTIVATION.

RL + EMPOWERMENT + MINECRAFT-ISH WORLD. POOR QUALITY

MOAR-15 GRADIENT DESCENT FOR GENERAL REINFORCEMENT LEARNING

IS POLICY GRADIENT. POLICY AND VALUE FCN TOGETHER. GENERAL FORMULATION. VAPS ALGO

MOAR-16 DETERMINISTIC POLICY GRADIENT ALGORITHMS

IS EXPECTED GRADIENT OF ACTION-VALUE FUNCTION. DETERMINISTIC POLICIES MORE EFFICIENT THAN STOCHASTIC PG.
OFF POLICY ACTOR CRITIC. TARGET: DETERMINISTIC BEHAVIOR: EXPLORATORY

MOAR-20 HOW CAN WE DEFINE INTRINSIC MOTIVATION?

SURVEY. INFORMATION THEORY PERSPECTIVE, KNOWLEDGE BASED MODELS, COMPETENCE, MORPHOLOGICAL

MOAR-24 POLICY GRADIENT METHODS FOR RL WITH FUNCTION APPROXIMATION

POLICY IS A FUNCTION → IE A ANN

MOAR-27 REINFORCE

PROTO POLICY GRADIENT. 1992