Update ¡date start¿ - ¡date end¿

Mason Smith

January 8, 2023

Agenda

Introduction

Current Work:

- Improving game design to allow for consistent training
- Training stable CPT value functions that represent biased policies
- Redesigning world and retraining to ensure non-trivial differences
  - Do not complete opposite or identical realization of strategy
  - Want to achieve the objective (catch the target) with semi-similar success rate in different ways (trajectories)
- Testing effects of assuming different biases for H in simulation

Challenges:

- Was struggling with getting stuck in local optima due to sparse gains (catch at end of game) and frequent penalties (throughout the game)
- There is a sensitive and fine balance for the following hyper-parameters that induce interesting policies to contrast:
  - Admissible bounds for risk-sensitivity
  - World design and initial conditions
  - Learning hyper-parameters

Agenda

Issues with Current Game

- Algorithm gets stuck in local optima due to sparse gains (catch at end of game) and frequent penalties (throughout the game)
- Strong encouragement to wait out the rest of the game once a penalty is received (do not chase target anymore)
- Different worlds induced differences in bias policies that were either
  - too strong
    - risk-averse had 0% catch rate while risk-seeking had 100% catch rate
    - impossible to evaluate coordination since strategies were incompatible → 0% catch
    - produced trivial result since both of the incorrect assumptions had same outcome
  - too weak
    - both policies either had near 100% or both had 0% performance
    - solution is so obvious that CPT does not change it
    - again produced trivial result due similarity between strategies

Update Goals

- Produce two policies (averse and seeking) that
  - produce compatible strategies that achieve the objective $p(success) - \epsilon > 0$ when paired
  - produce sufficiently unique joint-behavior when mis-matching policies compared to matching policies
  - produce sufficiently similar performance between matched-averse and matched-seeking policies s.t. comparison between is valid
- Modify
  - world configurations (initial positions and penalty positions)
  - global game rules and game hyper-parameters

List of Updates

- Redesigned world initial states and penalty locations
  - remove low-effect and difficult to train worlds
- Reward for catching target $r(catch) = 20 \rightarrow 25$
  - increasing window to receive positive reward $\sum r_t > 0$ after penalties and -1 turn reward are added
- Reward is now delivered as a single cumulative reward r_ζ at the end of the game
  - previously provided reward at every time-step r_t
  - helps avoid getting stuck local optima since intermediate rewards were only penalties
  - evaluates reward on a trajectory-scope $r_\zeta(\mathbf{s}_T, \mathbf{a}_T)$
  - instead of action-scope $r_t(s_t, a_t)$ where $r_\zeta = \sum_{t \in T} r_t$
  - apply eligibility traces ($TD(\lambda)$) to account for increased sparsity
- Reward is now non-negative
  - the cumulative reward at the end of the game is $r_\zeta = max(r_\zeta, 0)$
  - new objective is to maximize your reward **upon catching the target**
  - doing really bad and not catching the stag are now equivalent
  - eliminates trivial policy of avoiding rewards by never moving

Final Worlds

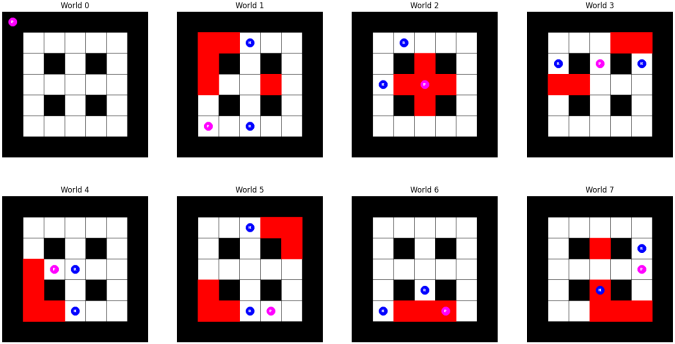


Figure 1: Updated World Designs

Agenda

Training Setup

- Implemented a independent joint-Q learning algorithm with directed exploration and Theory of Mind (ToM)
- Quantal Response Equilibrium (QRE) used as the equilibrium condition to solve games at every stage (ToM)
- Trained 3x policies π trained during self-play:
  - baseline/optimal (π_0)
  - risk-averse (π_A)
  - risk-seeking (π_S)
- Baseline policy π_0 was used as prior for biased policies π_A and π_S trained with cumulative prospect theory (CPT) agents
- Sophistication (level of recursion) was set to 3 in the QRE

Notation

- ego agent denoted by subscript $(\cdot)_k$ where $(\cdot)_{-k}$ represents the partner
- joint state $s \in S$ where $S = S_k \times S_{-k}$ given $\times$ denotes the Cartesian product
- joint action $a \in A$ where $A = A_k \times A_{-k}$ and a may be written as $\{a_k, a_{-k}\}$ for clarity
- ego stage reward r_t
- a policy π_k
  - always denotes choosing ego action a_k in s
  - samples a_k from *joint state-joint action* values $Q_k(s, a)$ given an est. $-k$ policy $\hat{\pi}_{-k}$
  - $Q_k(s, a)$ is reduced to *joint state-ego action* values $Q_k(s, a_k)$ by conditioning on $\hat{\pi}_{-k}$

    $$\text{s.t. } Q_k(s, a_k) = \mathbb{E}\big[Q_k(s, a | a = \{a_k, a_{-k} = \hat{\pi}_{-k}(s)\})\big] \; \forall a_k \in A_k$$
  - a_k is then drawn from $Q_k(s, a_k)$ according to a nominal Boltzmann distribution.
  - for brevity this reduction will be implied and we will write $Q_k(s, \{a_k, \hat{\pi}_{-k}(s)\})$ to denote the full expression $\mathbb{E}\big[Q_k(s, a | a = \{a_k, a_{-k} = \hat{\pi}_{-k}(s)\})\big] \; \forall a_k \in A_k$

Algorithm

Joint-$TD(\lambda)$

Initialize $Q_k(s, a)$ arbitrarily for all s, a
foreach *episode* **do**
  Initialize s and $e_k(s, a) = \mathbf{0}$
  foreach *step of episode* **do**
    $a_k \leftarrow$ ego action given by $\pi_k(s|\hat{\pi}_{-k}(s))$
    Take action a_k, observe joint action a, ego reward r_k, and next state s'
    $\delta \leftarrow r_k + \gamma \max_{a'_k} Q_k(s', \{a'_k, \hat{\pi}_{-k}(s')\}) - Q_k(s, a)$
    $e_k(s, a) \leftarrow e_k(s, a) + 1$
    foreach $s \times a$ **do**
      $Q_k(s, a) \leftarrow Q_k(s, a) + \alpha \delta e_k(s, a)$
      $e_k(s, a) \leftarrow \gamma \lambda e_k(s, a)$
    end foreach
    $s, a \leftarrow s', a'$
    Until s is terminal ;
  end foreach

Area Under the Indifference Curve (AUIC)

- area under the indifference curve (AUIC) is an expression of preference for accepting or rejecting a gamble over actions with certain outcomes in terms of probabilities $p(accept)$
- AUIC is evaluated over the space of feasible rewards $\mathbf{R}$ found in the game
- We define binomial-choices (a_1, a_2) with outcomes sampling from $\mathbf{R}$ s.t. $\mathbf{R}_1, \mathbf{R}_2 = \mathbf{R}$
- The outcomes of each choice are then:
  - a_1 containing one certain outcome
    - with possible rewards $\mathbf{R}_1 = \{r_1 - 0.5 * r_\rho \; \forall \, r_1 \in \mathbf{R}_1\}$
  - a_2 containing two uncertain outcomes (with and without a penalty r_ρ)
    - with possible rewards $\mathbf{R}_2 = \{[r_2, (r_2 - r_\rho)] \; \forall \, r_2 \in \mathbf{R}_2\}$
    - with probabilities $p = [(1 - p_\rho), p_\rho]$ for each outcome occurring
- *Indifference Curve*:
  - a continuous curve through the 2D reward space $(\mathbf{R}_1 \times \mathbf{R}_2)$
  - occurs when no preference is expressed s.t. $p(accept) = 1 - p(accept)$

Area Under the Indifference Curve (AUIC)

- $p(accept) = p(a_2)$ then implies risk-sensitivity where
  - An optimal agent expresses no preference (indifferent) given $r_1 = r_2 \; \forall \; r_1, r_2 \in \mathbf{R}$
  - Preferences become more complex as we apply CPT transformation $\mathbb{C}[\cdot]$
- AUIC will be calculated as follows:
  - Expresses the cumulative (mean) probability of $p(accept)$ across a symmetrical space of rewards transformed by CPT
  - Centered around 0 for legibility s.t. AUIC$\in (-0.5, 0.5)$
  - AUIC $= \frac{1}{|\mathbf{R}_1 \times \mathbf{R}_2|} \sum_{r_1, r_2 \in \mathbf{R}_1, \mathbf{R}_2} p(accept|\mathbb{C}[r_1, r_2]) - 0.5$
- The value for AUIC can then be interpreted as follows:
  - AUIC $+ p_\epsilon < 0$: the agent cumulatively prefers rejecting the gamble and is risk-averse
  - AUIC $- p_\epsilon > 0$: the agent cumulatively prefers accepting the gamble and is risk-seeking
  - $|\text{AUIC}| < p_\epsilon$: the agent agent has week cumulative preferences and is risk-insensitive
  - AUIC $= 0$: the agent has no cumulative preferences and is optimal
  - where $p_\epsilon = 0.1$ is a threshold defining what we consider

Area Under the Indifference Curve (AUIC)

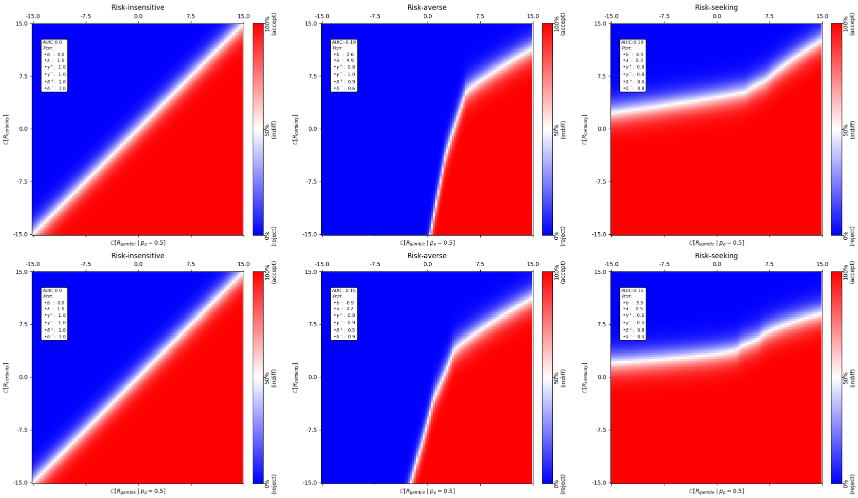


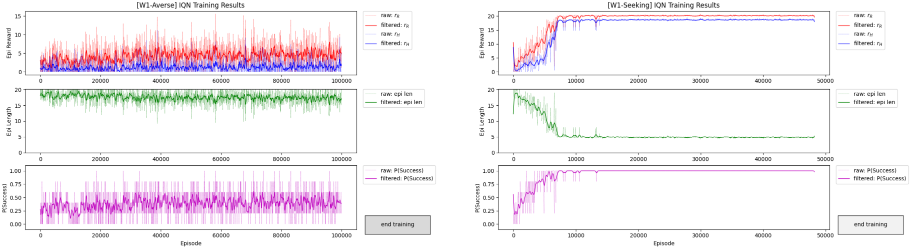
Figure 2: AUIC Samples

Creating Biased Policies

- CPT parameters $\mathcal{P}_{CPT}$ were stochastically perturbed while training biased policies
  - $\mathcal{P}_{CPT}$ were sampled in batches every 200 episodes
  - $\mathcal{P}_{CPT}$ were sampled from feasible bounds based on behavioral research [CITE]
  - $\mathcal{P}_{CPT}$ were attributed to averse or seeking behavior based on the AUIC
- $\mathcal{P}_{CPT}$ is continuously sampled until intended risk-sensitivity (AUIC) is met

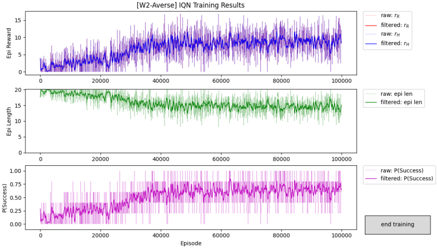
Convergence Expectations

- "Optimal strategy" is arbitrary between different bias conditions
- Different bias conditions induce different environment and therefore different policy
- Convergence conditions and final policy performance is not shared between bias conditions
- Seeking and baseline strategies may be similar due to world conditions
- It is somewhat hard to tell if a policy has converged for the averse condition π_A
    - Obvious convergence is not present.
    - averse induces higher penalties and less value in entering penalty states to chase target
    - Convergence often not evident from rewards, episode length, or probability of catching the target + MARL environments can be non-stationary
    - instead, relies on several iterations with varying learning parameters converging to similar results
    - had to update worlds[1] to balance between the risk of entering

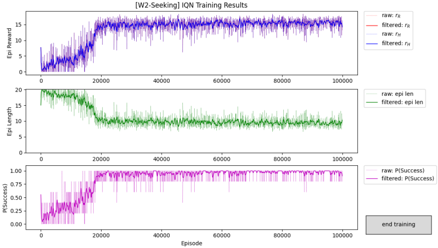


(a) Risk-Averse (b) Risk-Seeking

Figure 3: World 1 Training Results



(a) Risk-Averse

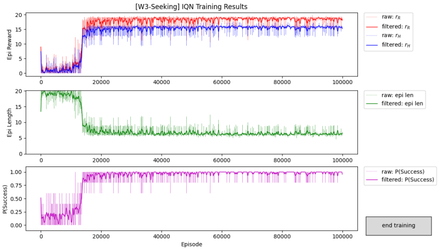


(b) Risk-Seeking

Figure 4: World 2 Training Results

Training Results (World 3)

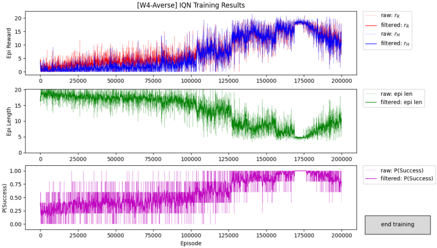
(a) Risk-Averse



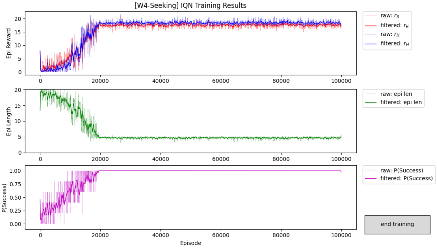
(b) Risk-Seeking

Figure 5: World 3 Training Results

Training Results (World 4)



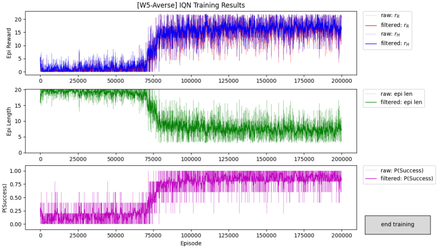
(a) Risk-Averse



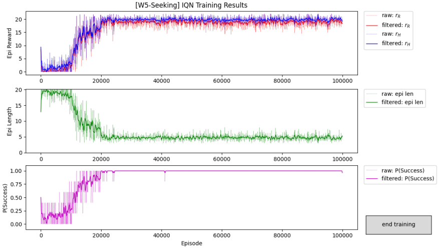
(b) Risk-Seeking

Figure 6: World 4 Training Results

Training Results (World 5)



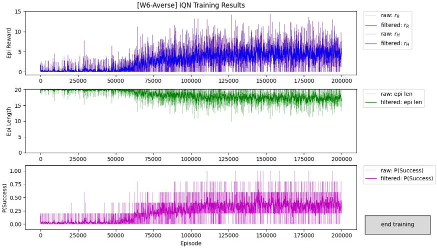
(a) Risk-Averse



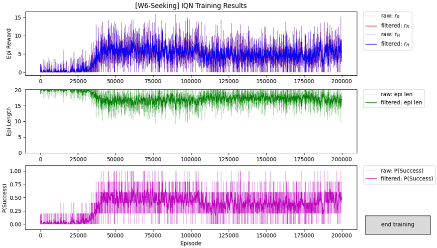
(b) Risk-Seeking

Figure 7: World 5 Training Results

Training Results (World 6)



(a) Risk-Averse



(b) Risk-Seeking

Figure 8: World 6 Training Results

Discussion

- Policies are generally noisy due not non-stationary and rationality constant $= 1$
- Equilibrium would be less stochastic with higher rationalities
- Noise in final result pseudo-required to avoid the previously mentioned all or nothing problem (e.i. there exists dis-coordination and vulnerability to partner uncertainty)
- Baseline (Optimal) policies are often similar to Risk-Seeking policies since the game is designed for the agents to succeed
  - Rushing through penalties is often a good strategy
  - Susceptible to partner and target stochasticity
- May make minor attempts to improve policies in future but this is good for now

Agenda

Setup

Goal:

- Evaluate how assumptions of H's risk-sensitivity effect team performance
- Provide validation that there are differing or conflicting optimal policies based on risk-sensitivity

Experimental Conditions:

- We manipulate
  - what R assumes H's policy to be ($\hat{\pi}_H$)
  - what H's policy actually is (π_H)
- The experimental condition is then written as $\mathcal{C} = \{\hat{\pi}_H, \pi_H\}$
- Substituting in our bias policies that we trained we get four conditions:
  - Assume-Averse + Is-Averse: $\{\hat{\pi}_A, \pi_A\}$ (Correct Assumption)
  - Assume-Seeking + Is-Averse: $\{\hat{\pi}_S, \pi_A\}$ (Incorrect Assumption)
  - Assume-Averse + Is-Seeking: $\{\hat{\pi}_A, \pi_S\}$ (Incorrect Assumption)
  - Assume-Seeking + Is-Seeking: $\{\hat{\pi}_S, \pi_S\}$ (Correct Assumption)
  - *can include optimal vs X conditions but small effect is present*

Analysis

Approach:

- We **only compare between R's assumption** and within H's actually policy
- H's actually policy directly effects game performance and invalidates some evaluation metrics
- R's policy will be $\pi_R = \pi_0$ conditioned on $\hat{\pi}_H$ using QRE
- H will assume R uses H's true policy $\hat{\pi}_R = \pi_H$
- Run simulated game 1000x for each of the 4 conditions composed

Metrics:

- Each agent's reward and the team (mean) reward between assumptions
- Episode length and probability of catching the target between assumptions
- Number of penalty states each agent during an average game between assumptions
- Mean probability of partner's action in ego's mental model $p(a_{-k}|\hat{\pi}_{-k})$

Hypothesis

- When R assumes wrong ($\hat{\pi}_H \neq \pi_H$) for both $\pi_H \in (\pi_A, \pi_S)$:

H1.1: Each agent's and team reward will decrease
H1.2: Episode length will increase
H1.3: Probability of catching target will decrease
H1.3: Number of penalty states entered will increase
H1.4: Both agents will not be able to predict each other's actions well (small $p(a_{-k}|\hat{\pi}_{-k})$)

- When H is averse ($\pi_H = \pi_A$) instead of seeking:

H2.1: Magnitude of performance losses will be less significant
H2.2: Performance will be worse than if $\pi_H = \pi_S$

- Misc. Hypothesises

H3: Only minor changes in terminal state location will occur when agents succeed (e.i. objective remains the same but joint-trajectory changes).

Analysis

Metrics:

-
- * on top of bars and in key indicate correct assumption made

Results

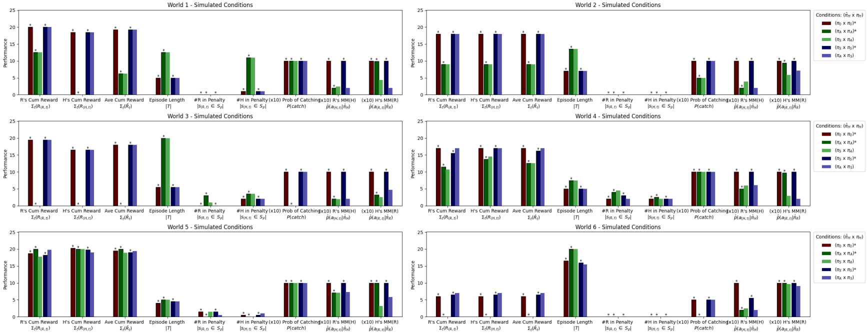


Figure 9: Evaluation of simulated conditions per world

Results

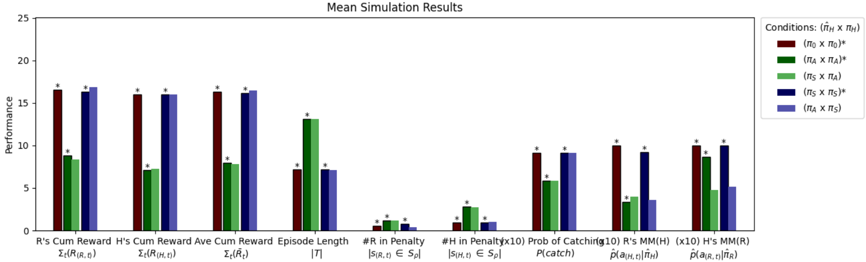


Figure 10: Evaluation of simulated conditions summary

Discussion

- ADD DISCUSSION

Agenda

2-Week Sprint Goals

Goal:

- Description

Long-Term Goals

Goal:

- Description