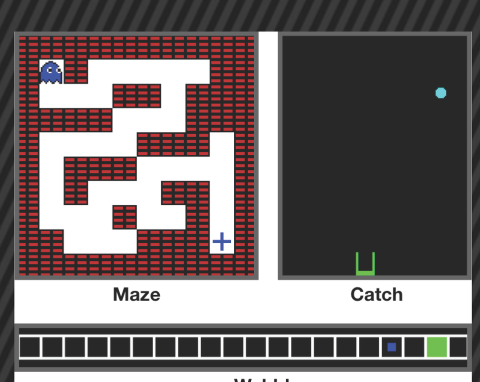
Playing Games with Implicit Human Feedback

Mohit Agarwal*, Duo Xu*, Faramarz Fekri, Raghupathy Sivakumar
Electrical and Computer Engineering, Georgia Institute of Technology

Overview

- Human feedback can significantly accelerate the Reinforcement Learning (RL) algorithms in end-user applications
  - Human ratings and rankings [El Asri et al. 2016]
  - Learning from Demonstrations [Ng, Harada, and Russell 1999]
- The issues with explicit human feedback
  - Severely burdens the human involved in the loop
  - Explicitly requires the humans to take actions
  - Difficult (or impossible) in some situations like driving (or disable user)
- **Implicit human feedback:** Humans' intrinsic reactions as implicit (and natural) feedback through Electroencephalography (EEG) in the form of error-related potentials (ErrPs)
  - Inspired by a high-level error-processing system in humans that generates error-related potential (ErrP) [Scheffers et al., 1996]
  - When a human recognizes an error made by an agent, elicited ErrP can inform about the sub-optimality of executed action in the given state
- Widens the applicability of RL-human interactive systems
  - Feedback is direct and fast while being natural and easy for humans
  - Avoids unwanted situations with increased latency (explicit human feedback) in real-world environments

System Setup and Data Collection



- Developed three discrete-grid based environments in OpenAI Gym Atari framework
  - Wobble, Catch and Maze
  - https://github.com/meagmohit/gym-maze
- Experimental Protocol (approved by IRB)
  - Machine agent plays a computer game while a human silently observes
  - Agent took action every 1.5 seconds
- Hardware: OpenBCI Cyton w/ BIOPAC CAP
- Software: OpenViBE platform + OpenAI Gym
- Recruited 5 subjects (mean age = 26.8)

Naïve Approach to Integrate Implicit Human Feedback

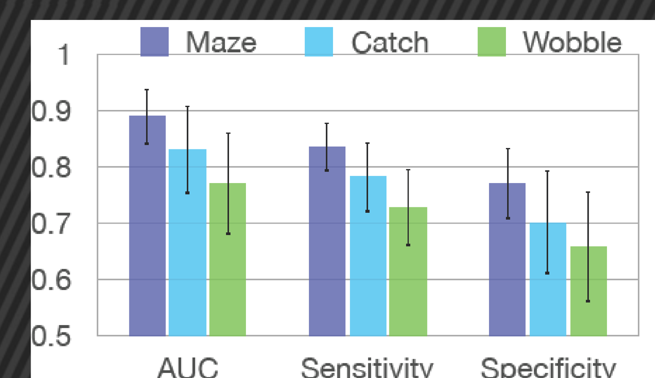
Obtaining the Implicit Human Feedback

Riemannian Geometry based ErrP decoding [Barachant et al., 2014]
- State-of-the-art algorithm for decoding event-related potentials
- Binary classification problem for ErrP labels
- Performance using 10-fold cross-validation
  - AUC of 0.89 for Maze, 0.83 for Catch and 0.77 for Wobble
  - Over 80% sensitivity for Maze

Algorithm 1: Riemannian Geometry based ErrP classification algorithm

Input : raw EEG signals EEG
1. Pre-process raw EEG signals ;
2. Spatial Filtering: xDAWN Spatial Filter ($nfilter$) ;
3. Electrode Selection: ElectrodeSelect ($nelec$, metric='riemann') ;
4. Tangent Space Projection : TangentSpace(metric = "logeuclid") Normalize using L1 norm ;
5. Regression: ElasticNet ;
6. Select decision threshold by maximizing accuracy

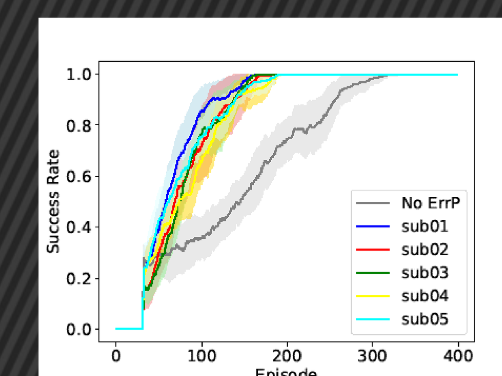
Algorithm: Decoding ErrPs



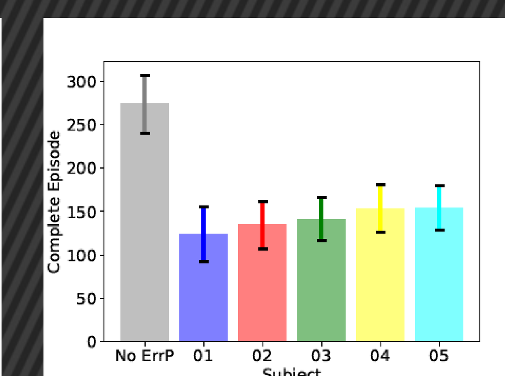
Performance: Decoding ErrPs

Integrating the Implicit Human Feedback

Reward shaping with full access method
- Human feedback is obtained on every state-action pair
- Time-intensive and not practically feasible
- Performance Evaluation
  - Success rate: ratio of successful plays in the last 32 episodes
  - Achieves a training acceleration of 2.25x



Learning Curve

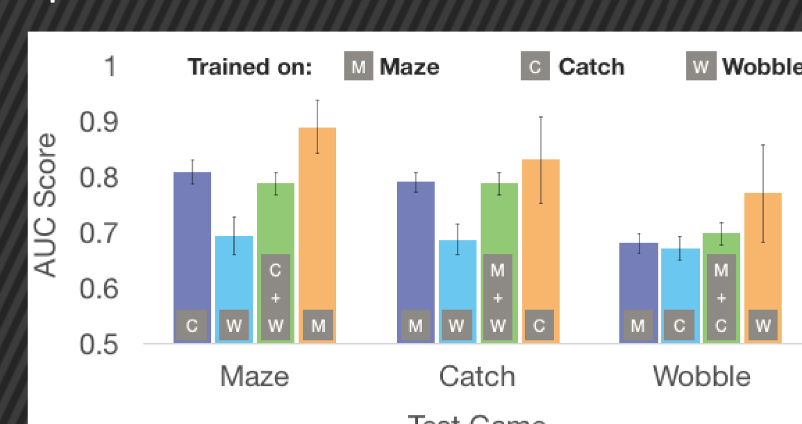


Complete Episode

Towards Practical Integration of Implicit Human Feedback

Zero-shot learning of ErrPs

- Definition of error-potentials can be learned in a zero-shot manner
- Experimentally validate that ErrPs can be learned on one environment, and the decoder is used as-is for novel and unseen environments
- Performance:
  - AUC of 0.9078 (test: Maze, train: Catch)
  - Captures more than 80% of variability compared to 10-fold CV



Performance: zero-shot learning over all game combinations compared with 10-fold CV

Learning from Imperfect Demonstrations

- Implicit human feedback is required on initially given trajectories
- An auxiliary reward function is learned based on the labeled trajectories prior to RL training
- During the RL training, the learned reward function acts as a proxy for the human feedback
- Queries are made initially on a subset of state-action space
  - Reduces the total number of queries and hence, cognitive load on humans
- Performance:
  - Proposed approach makes 75.56% fewer queries as compared to *full access*
  - Achieves 2.25x acceleration averaged over 5 subjects

Algorithm 2: Leraning from Imperfect Demonstration by Human ErrP

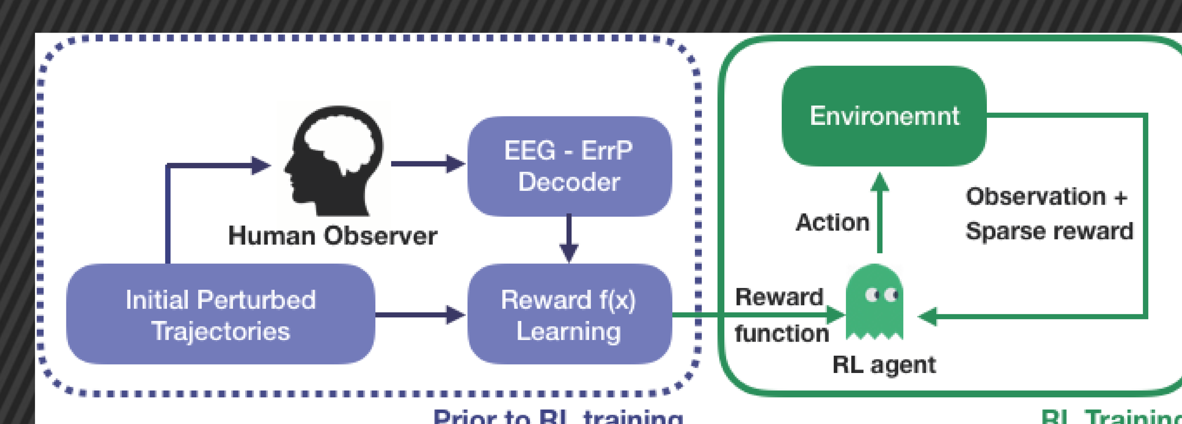
Input : Trajectories Given Initially
1. Conduct EEG Experiments to label the correctness along trajectories by human ErrP ;
2. With ErrP data collected in experiments, use Algorithm 1 to decode ErrP labels, i.e., ErrP$(\cdot,\cdot)$;
3. Learn the preliminary Q function $Q(\cdot,\cdot)$ by optimizing

$$J_1(Q) := \sum_{(s,a)\in\mathcal{D}} \pi_Q(a|s)(1-\mathrm{ErrP}(s,a)) + (1-\pi_Q(a|s))\mathrm{ErrP}(s,a) \quad (2)$$

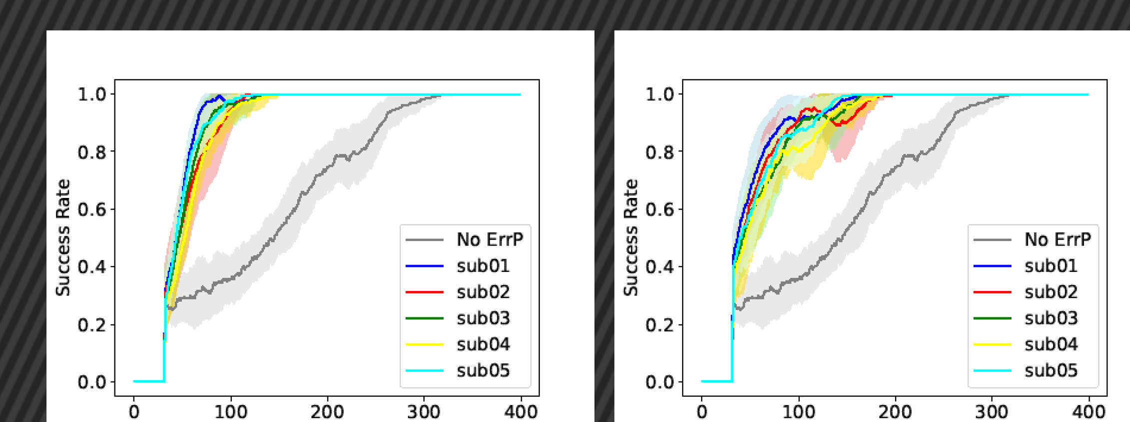
and learn the reward shaping function $t(\cdot)$ by minimizing

$$J_2(t) := \sum_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\in\mathcal{D}\cup\mathcal{D}_R} l(Q(\boldsymbol{s},\boldsymbol{a}) - t(\boldsymbol{s}) - \gamma \max_{\boldsymbol{a}'\in\mathcal{A}}(Q(\boldsymbol{s}',\boldsymbol{a}') - t(\boldsymbol{s}'))) \quad (3)$$

Then we have $Q_B(\boldsymbol{s},\boldsymbol{a}) := Q(\boldsymbol{s},\boldsymbol{a}) - t(\boldsymbol{s})$;
4. Transmit the auxiliary reward function $r'(\boldsymbol{s},\boldsymbol{a}) := Q_B(\boldsymbol{s},\boldsymbol{a}) - \gamma \max_{\boldsymbol{a}'\in\mathcal{A}} Q_B(\boldsymbol{s}',\boldsymbol{a}')$ to the RL agent ;
5. RL agent starts to solve the problem with reward function $r + r'$ by any RL algorithm.



Proposed Framework



Learning Curves
10 trajectories 20 trajectories

Future Work

- Scalability over environments with larger state-spaces
  - Extending the scope of zero-shot learning beyond discrete-grid navigational games
- Integrating human feedback
  - Preserving policy optimality with reward shaping
  - Considering other approaches e.g., policy shaping, IRL etc.

Acknowledgements

- NSF: CPS-1837369
- Wayne J. Holman Endowed Chair