Motivation

- Weak supervision signals are everywhere (e.g., noisy reward or low-quality demonstrations) in sequential learning problems!
- Existing reinforcement learning (RL) and behavioral cloning (BC) algorithms rely on high-quality supervision signals, resulting in unstable or sub-optimal results when meeting weak supervisions.
- Some previous works have explored these topics separately in their specific domains. However, there lacks a unified solution for robust policy learning in imperfect situations.

Policy Learning from Weak Supervision

- We use $Y$ to denote a weak supervision. It could be noisy reward $\tilde{r}$ for RL or noisy action $\tilde{a}$ from an imperfect expert policy $\tilde{\pi}$. For each weakly supervised state-action pair $((s_i, a_i), \tilde{Y}_i)$, we randomly sample a state-action pair $(s_j, a_j), j \neq i$, as well as another supervision signal $\tilde{Y}_k, k \neq i, j$ from a different state-action pair. Then we evaluate $((s_i, a_i), \tilde{Y}_i)$ according to the following:

CA with Weak Supervision: $Eva_i((s_i, a_i), \tilde{Y}_i) = Eva_i((s_i, a_i), \tilde{Y}_i)$

- Intuition: (a) the first term above encourages an “agreement” with the weak supervision (b) the second term punishes a “blind” agreement that happens when the agent’s policy always matches with the weak supervision (c) the third term is a reward that is evaluated against another stronger signal $\tilde{Y}_k$, $k \neq i, j$. A simple way to evaluate the predicted action given the expert action $\tilde{a}_i$: $Eva_{BC}((s_i, a_i), \tilde{a}_i) = \log \pi(s_i | a_i)$

Goal: maximize $\mathbb{E}[\tau] = \mathbb{E}[Eva_i((s_i, a_i), \tilde{Y}_i)]$, where $\tau$ is the trajectories collected by learned policy $\pi$ or the demonstration dataset.

Solution: Correlated Agreement with Weak supervision

For each weakly supervised state-action pair $((s_i, a_i), \tilde{Y}_i)$, we randomly sample a state-action pair $(s_j, a_j), j \neq i$, as well as another supervision signal $\tilde{Y}_k, k \neq i, j$ from a different state-action pair. Then we evaluate $((s_i, a_i), \tilde{Y}_i)$ according to the following:

CA with Weak Supervision: $Eva_i((s_i, a_i), \tilde{Y}_i) - Eva_i((s_j, a_j), \tilde{Y}_k)$

- Intuition: (a) the first term above encourages an “agreement” with the weak supervision (b) the second term punishes a “blind” agreement that happens when the agent’s policy always matches with the weak supervision even on randomly paired traces.

Why Peer Reward Works?

- Hypothesis 1: PeerRL reduces the bias (while with larger variance like Wang et al., 2020)

  noisy reward: $E[r] = \eta - (1 - \epsilon) e + \epsilon$  
  peer reward: $E[p_{Peer}] = \eta - (1 - \epsilon) e + \epsilon$

- Hypothesis 2: PeerRL helps break ties

  1. “tie” states indicate that the rewards for different states are the same - unstable and uncertain
  2. randomness in discretization model thus breaking ties - more informative for optimization

Conclusion

- We formulated “weakly supervised policy learning” to unify a series of RL/BC problems with low-quality supervision signals.
- A theoretical principled framework PeerPL that builds on evaluating a learning policy’s correlated agreements with the weak supervisions.

Past Works:
1. Reinforcement Learning with Perturbed Reward. Wang et al., AAAI 2020
2. Peer Loss Functions: Learning from Noisy Labels without Knowing Noise. Liu et al., ICML 2020