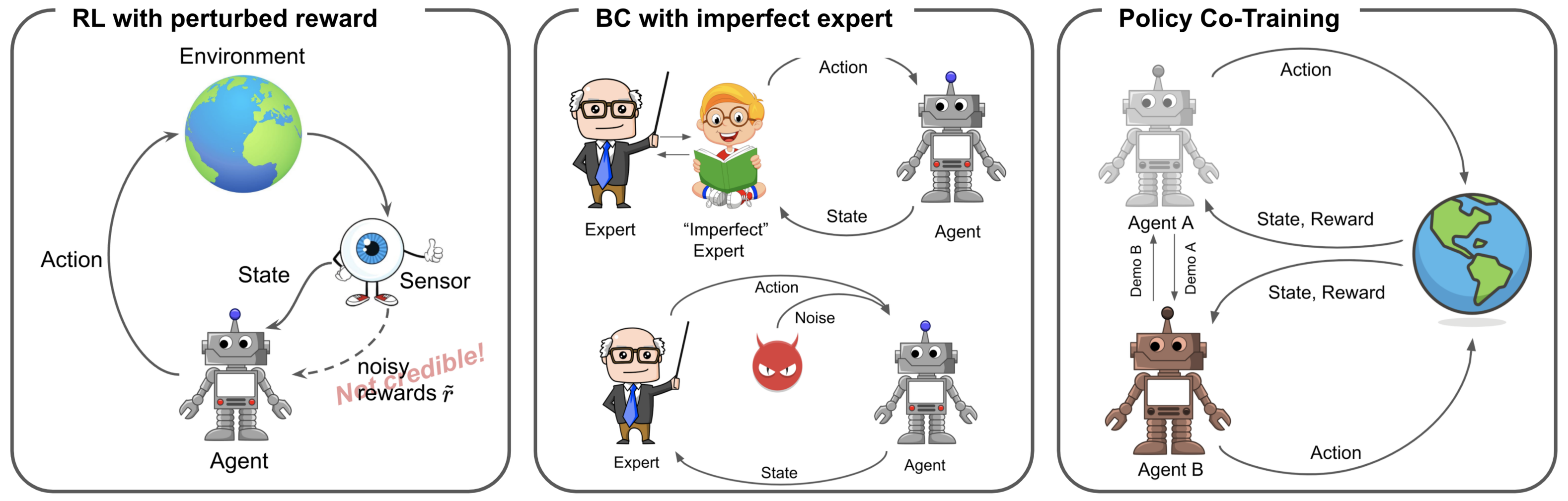
Policy Learning Using Weak Supervision

Jingkang Wang*[1,2], Hongyi Guo*[3], Zhaowei Zhu*[4], Yang Liu[4]
[1]University of Toronto, [2]Vector Institute, [3]Northwestern University, [4]University of California, Santa Cruz
wangjk@cs.toronto.edu, hongyiguo2025@u.northwestern.edu, {zwzhu,yangliu}@ucsc.edu,
Code available at: https://github.com/wangjksjtu/PeerPL

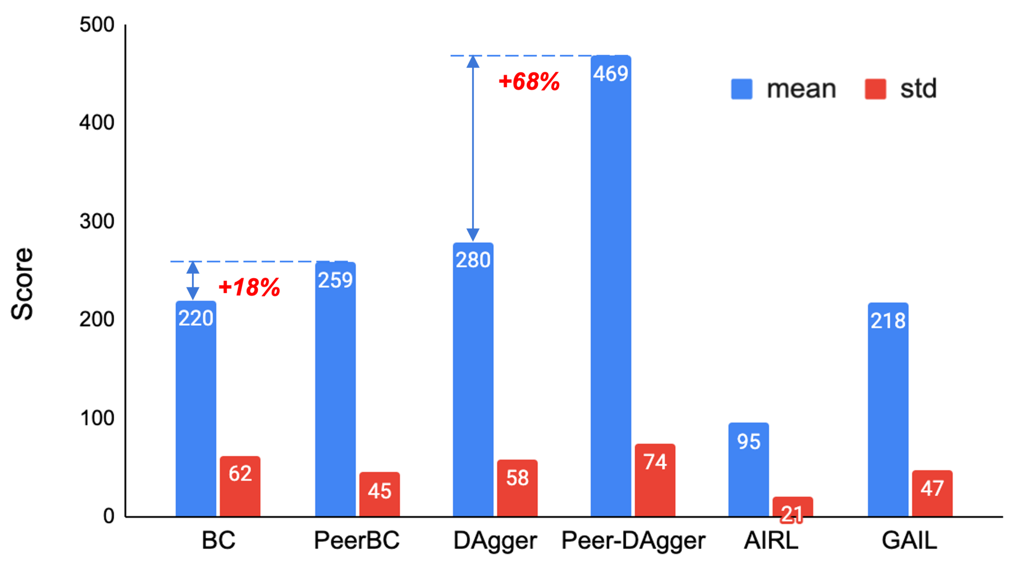
Motivation

- Weak supervision signals are everywhere in sequential learning problems (e.g., noisy reward or low-quality demonstrations)!



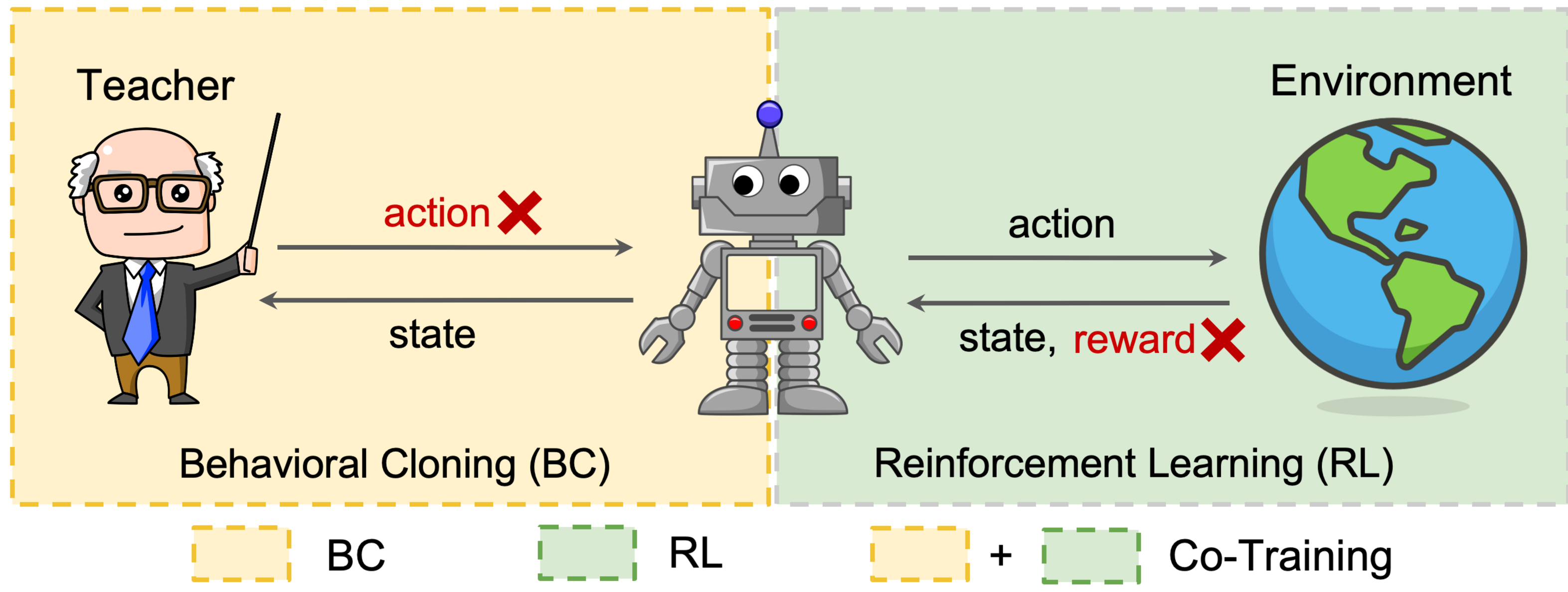
Weak Supervision:
- **RL:** The reward may be collected through sensors thus noisy
- **IL:** The demonstrations by an expert are often imperfect due to limited resources

- Most 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.



Policy Learning from Weak Supervision

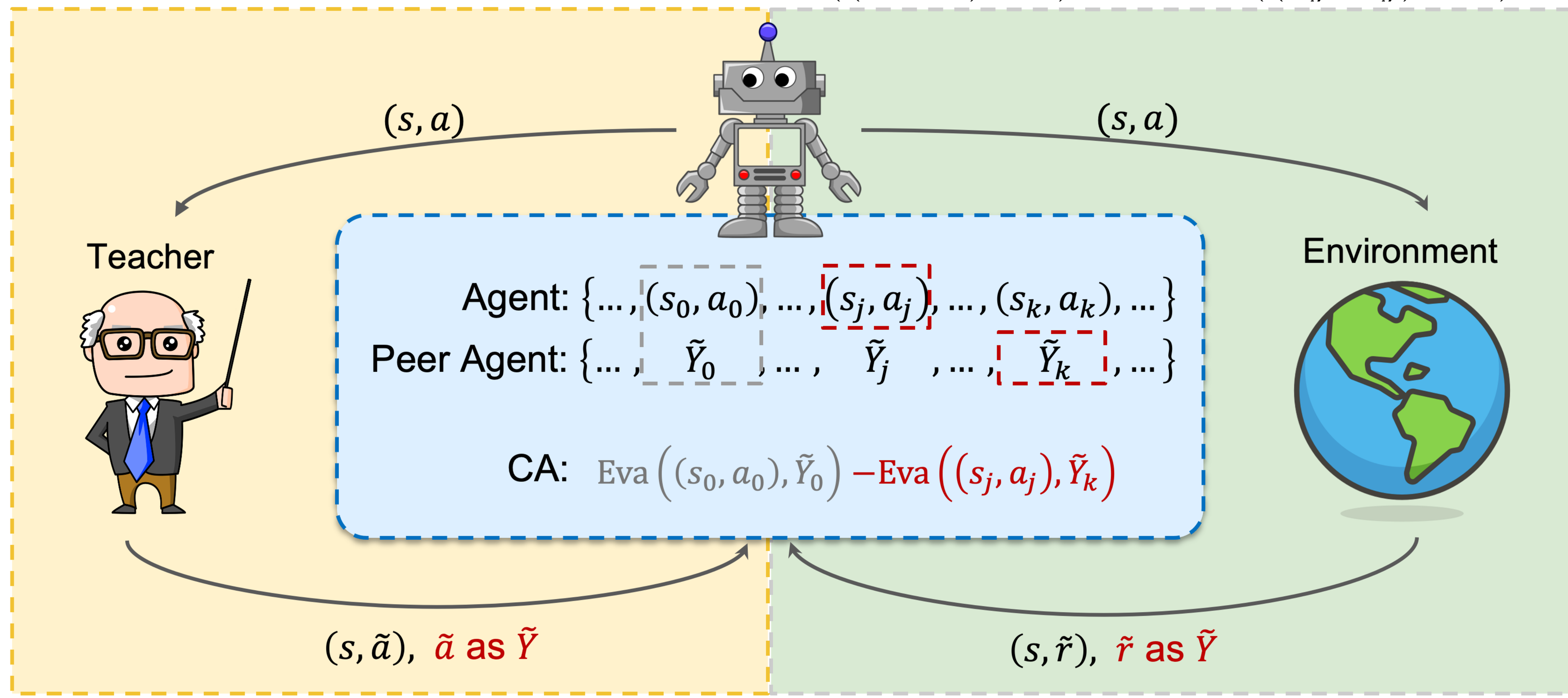
- We use $\tilde{Y}$ to denote a weak supervision. It could be noisy reward $\tilde{r}$ for RL or noisy action $\tilde{a}$ from an imperfect expert policy $\tilde{\pi}_E$ for BC.
- **Assumptions:**
  1. We consider a discrete noise model where the noise corruption can be characterized via a unknown confusion matrix: - $\mathbf{C}^{\mathrm{RL}}_{|\mathcal{R}|\times|\mathcal{R}|}$ or $\mathbf{C}^{\mathrm{BC}}_{|\mathcal{A}|\times|\mathcal{A}|}$.
  2. Only deterministic reward or expert policy is considered as it is hard to distinguish a clean case with noisy one without addition knowledge.
- **Objective**: Learning the optimal policy π^* with only a weak supervision sequence denoted as $\{(s_t, a_t), \tilde{Y}_t\}_{t=1}^T$ (RL) or $\{(s_i, a_i), \tilde{Y}_i\}_{i=1}^N$ (BC).



PeerPL with Correlated Agreement

- **A unified evaluation function**: Eva$_\pi$ to evaluate a taken policy π at agent state (s_i, a_i) using the weak supervision $\tilde{Y}_i$.
  - **(RL)** instance-wise measure (negative loss): a function of the noisy reward $\tilde{r}$ received at (s_i, a_i): $\mathrm{Eva}^{\mathrm{RL}}_\pi((s,a),\tilde{r}) = -\ell(\pi,(s,a,\tilde{r}))$
  - **(BC)** loss to evaluate the predicted action given the expert action $\tilde{a}_i$: $\mathrm{Eva}^{\mathrm{BC}}_\pi((s,a),\tilde{a}) = \log \pi(\tilde{a}|s)$
- **Goal**: maximize $J(\pi) = \mathbb{E}_{(s,a)\sim\tau}[\mathrm{Eva}_\pi((s,a),\tilde{Y})]$, where τ is the trajectories collected by learned policy π 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 : $\mathrm{Eva}_\pi((s_i,a_i),\tilde{Y}_i) - \mathrm{Eva}_\pi((s_j,a_j),\tilde{Y}_k)$



- **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: $\mathbb{E}[\tilde{r}] = \eta \cdot \left(\mathbb{E}[r] + \frac{e_+}{1-e_- - e_+} r_- + \frac{e_-}{1-e_- - e_+} r_+\right)$

peer reward: $\mathbb{E}[\tilde{r}_{\mathrm{peer}}] = \eta \cdot (\mathbb{E}[r] - (1-p_{\mathrm{peer}}) r_- - p_{\mathrm{peer}} r_+)$

potentially much larger than $(1-p_{\mathrm{peer}})$ and p_{peer} in high noise regime!

- **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

2-state Markov process (no actions)

$r_1 \sim \mathrm{clamp}[\mathcal{N}(0.6,1), \min=0, \max=1]$

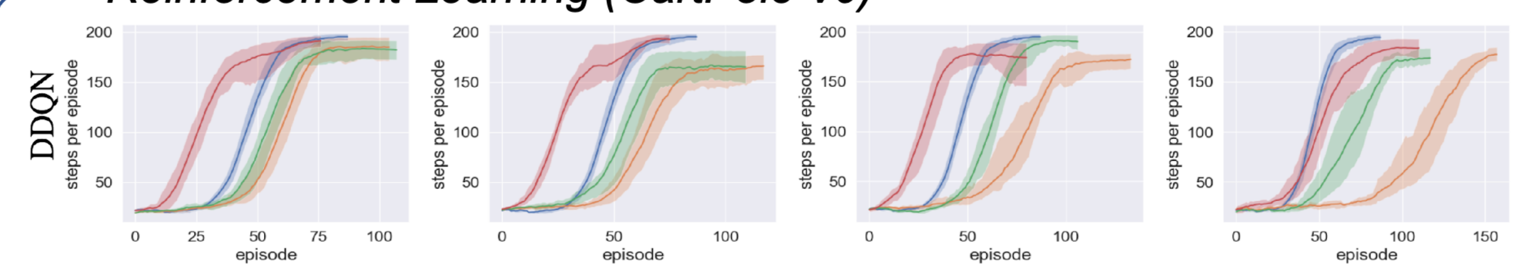
$r_2 \sim \mathrm{clamp}[\mathcal{N}(0.4,1), \min=0, \max=1]$

	Correct	Tie	Incorrect
Baseline	54.6%	5.6%	39.8%
PeerRL	58.0%	0.3%	41.7%

Tie breaking!

Experimental Results

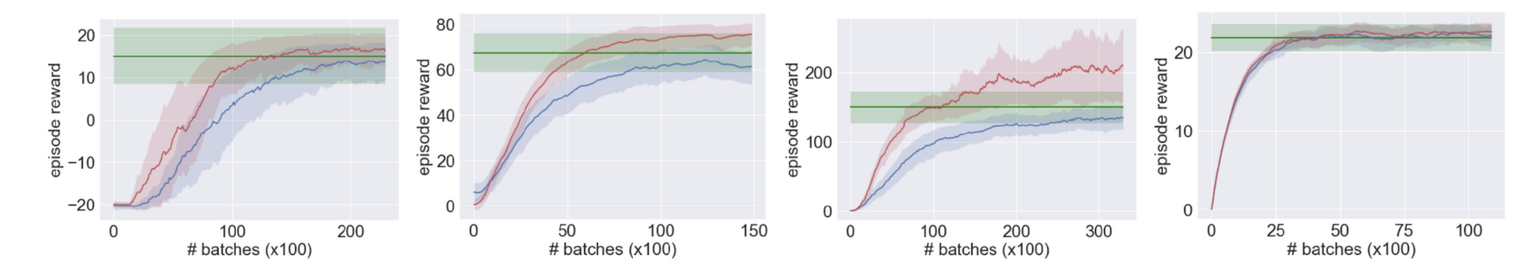
Reinforcement Learning (CartPole-v0)



(a) $e = 0.1$ (b) $e = 0.2$ (c) $e = 0.3$ (d) $e = 0.4$

True reward · Noisy reward · Surrogate reward (Wang et al, 2020) · PeerRL (ours)

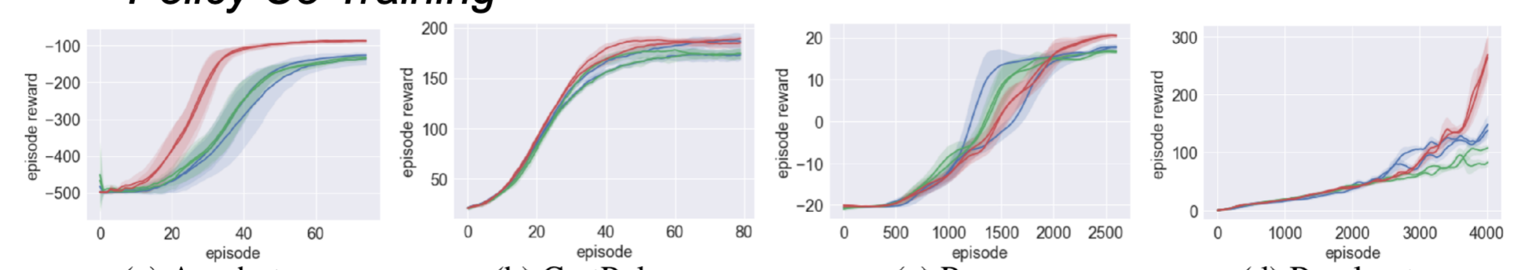
Behavioral Cloning



(a) Pong (b) Boxing (c) Enduro (d) Freeway

Standard BC · Weak expert · PeerBC (ours)

Policy Co-Training



(a) Acrobot (b) CartPole (c) Pong (d) Breakout

Single view · Co-Training (Song et al., 2019) · Peer Co-Training (ours)

Environment		Pong	Boxing	Enduro	Freeway	Lift (↑)
Expert		15.1 ± 6.6	67.5 ± 8.5	150.1 ± 23.0	21.9 ± 1.7	-
Standard BC		14.7 ± 3.2	56.2 ± 7.7	138.9 ± 14.1	22.0 ± 1.3	-6.6%
PeerBC	$\xi = 0.2$	$\mathbf{18.8 \pm 0.6}$	67.2 ± 8.4	177.9 ± 29.3	$\mathbf{22.5 \pm 0.6}$	$+11.3\%$
	$\xi = 0.5$	16.6 ± 4.0	$\mathbf{75.6 \pm 5.4}$	$\mathbf{230.9 \pm 73.0}$	22.4 ± 1.3	$\mathbf{+19.5\%}$
	$\xi = 1.0$	16.7 ± 4.3	69.7 ± 4.7	230.4 ± 61.6	8.9 ± 4.9	$+2.0\%$
Fully converged PPO		20.9 ± 0.3	89.3 ± 5.4	389.6 ± 216.9	33.3 ± 0.8	-

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 Rates. *Liu et al., ICML 2020.*