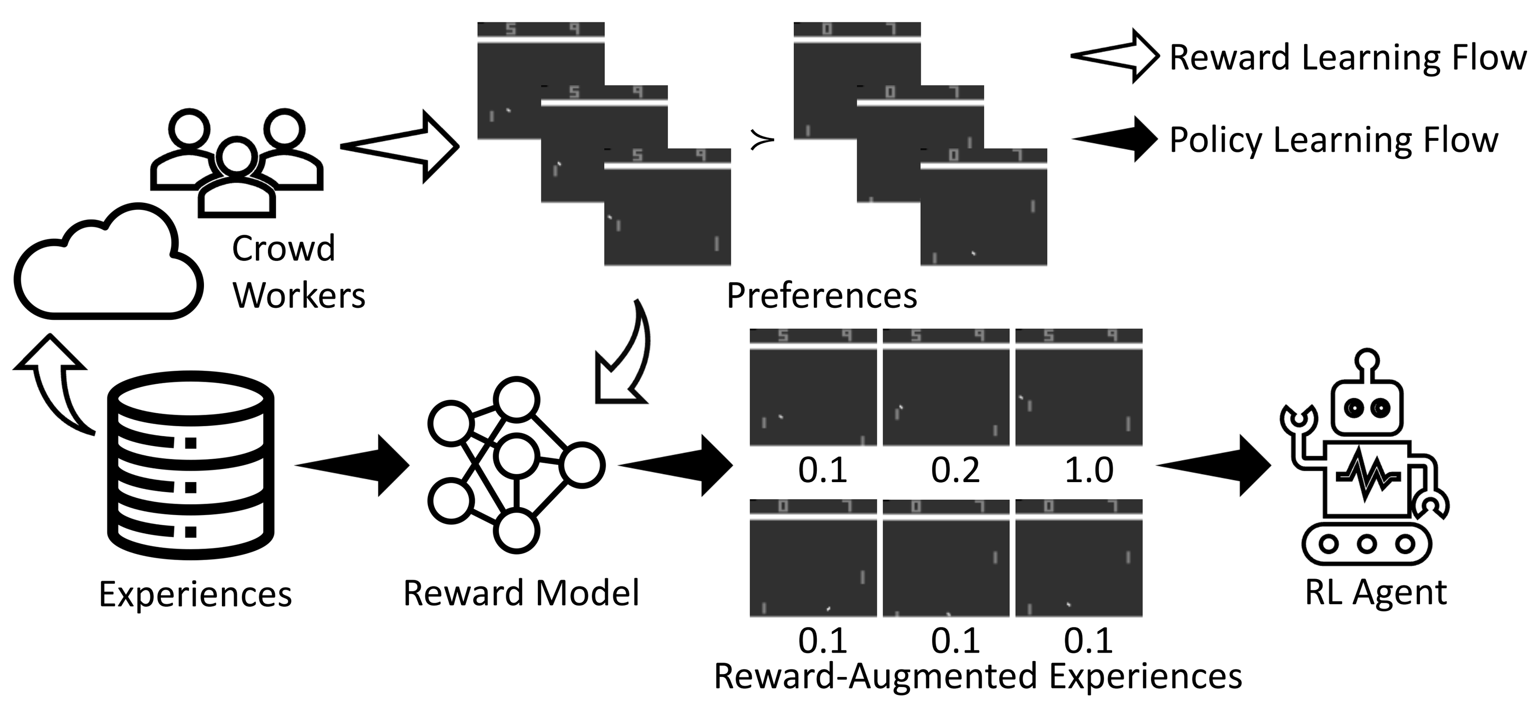
Batch Reinforcement Learning from Crowds

Guoxi Zhang, Hisashi Kashima

Graduate School of Informatics, Kyoto University

Motivation

- Batch reinforcement learning (RL) relies on rewards to refine policies. For tasks without reward signals, one may learn a reward function from human preferences over experiences.
- This study investigates how to learn a reward function from nonexpert annotators, which allows for leveraging crowdsourcing for batch RL. The main challenge is **denoising**, as nonexpert annotators make mistakes in preferences.



Model

This study proposes a probabilistic model named deep crowd-BT (DCBT) for learning a reward function from noisy preferences.

$\eta_{i,1}$ and $\eta_{i,2}$	The two trajectories in the i^{th} preference sample.
$R(s,a)$	The reward function to be learned.
$\sigma(\cdot)$	The sigmoid function $\sigma(x) = 1/(1+\exp(-x))$.
α_i	The reliability of the preference label in the i^{th} sample.
w_i	The ID of the annotator who labeled the i^{th} sample.
$\tilde{y}_i$	$\tilde{y}_i = 1$ if $\eta_{i,1} \succ \eta_{i,2}$, $\tilde{y}_i = 0.5$ if $\eta_{i,1} \approx \eta_{i,2}$, and $\tilde{y}_i = 0$ otherwise.

$$\mathrm{P}_{\mathrm{DCBT}}(\eta_{i,1} \succ \eta_{i,2}) = \boxed{\alpha_i}\mathrm{P}_{\mathrm{BT}}(\eta_{i,1} \succ \eta_{i,2}) + (1-\alpha_i)\left(1 - \mathrm{P}_{\mathrm{BT}}(\eta_{i,1} \succ \eta_{i,2})\right)$$

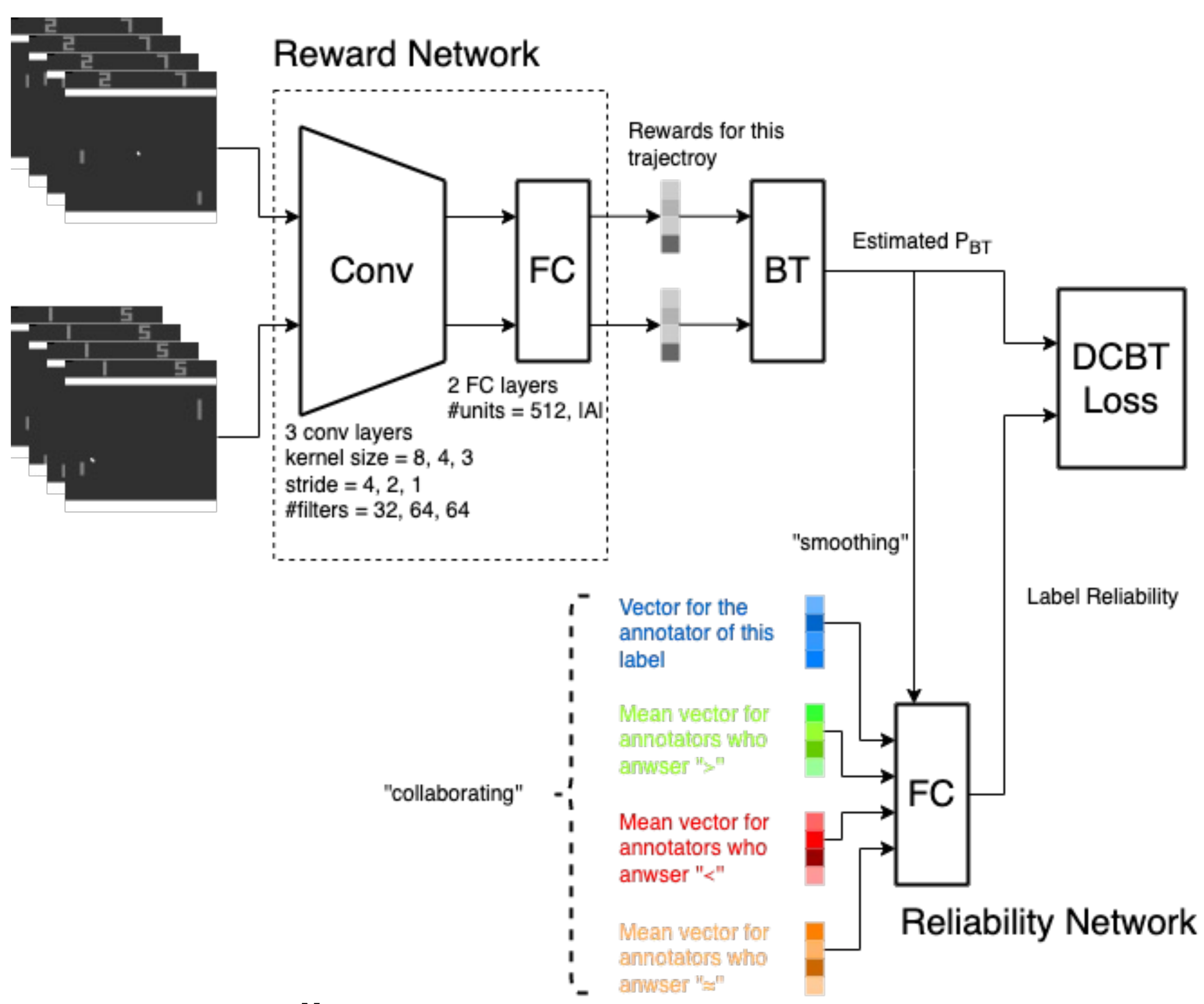
α_i depends on three factors:

- w_i
- $\mathrm{P}_{\mathrm{BT}}(\eta_{i,1} \succ \eta_{i,2})$
- Other labels for the same pair of trajectories and the corresponding annotators

$$\mathrm{P}_{\mathrm{BT}}(\eta_{i,1} \succ \eta_{i,2}) = \sigma\left(G(\eta_{i,1}) - G(\eta_{i,2})\right)$$

$$G(\eta) = \frac{1}{|\eta|}\sum_{(s,a)\in\eta} R(s,a)$$

Without noise, the trajectory with larger average reward is preferred.



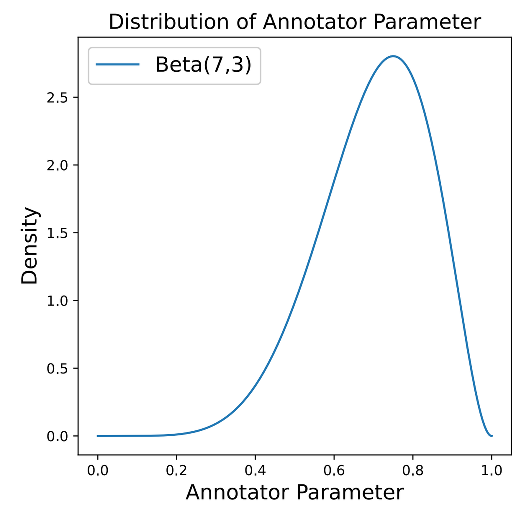
$$L_{\mathrm{DCBT}} = -\frac{1}{N}\sum_{i=1}^{N} \tilde{y}_i \log(\mathrm{P}_{\mathrm{DCBT}}) + (1-\tilde{y}_i)\log(1-\mathrm{P}_{\mathrm{DCBT}})$$

Learning

- Use ℓ_1 and ℓ_2 regularization, also regularize rewards toward zero by: $-\frac{1}{2N}\sum_i \sum_{k=1,2} \log(\sigma(G(\eta_{i,k}))) + \log(\sigma(-G(\eta_{i,k})))$.
- Initialize the reward network by fixing α_i to 0.99 and update the reward network using preferences.

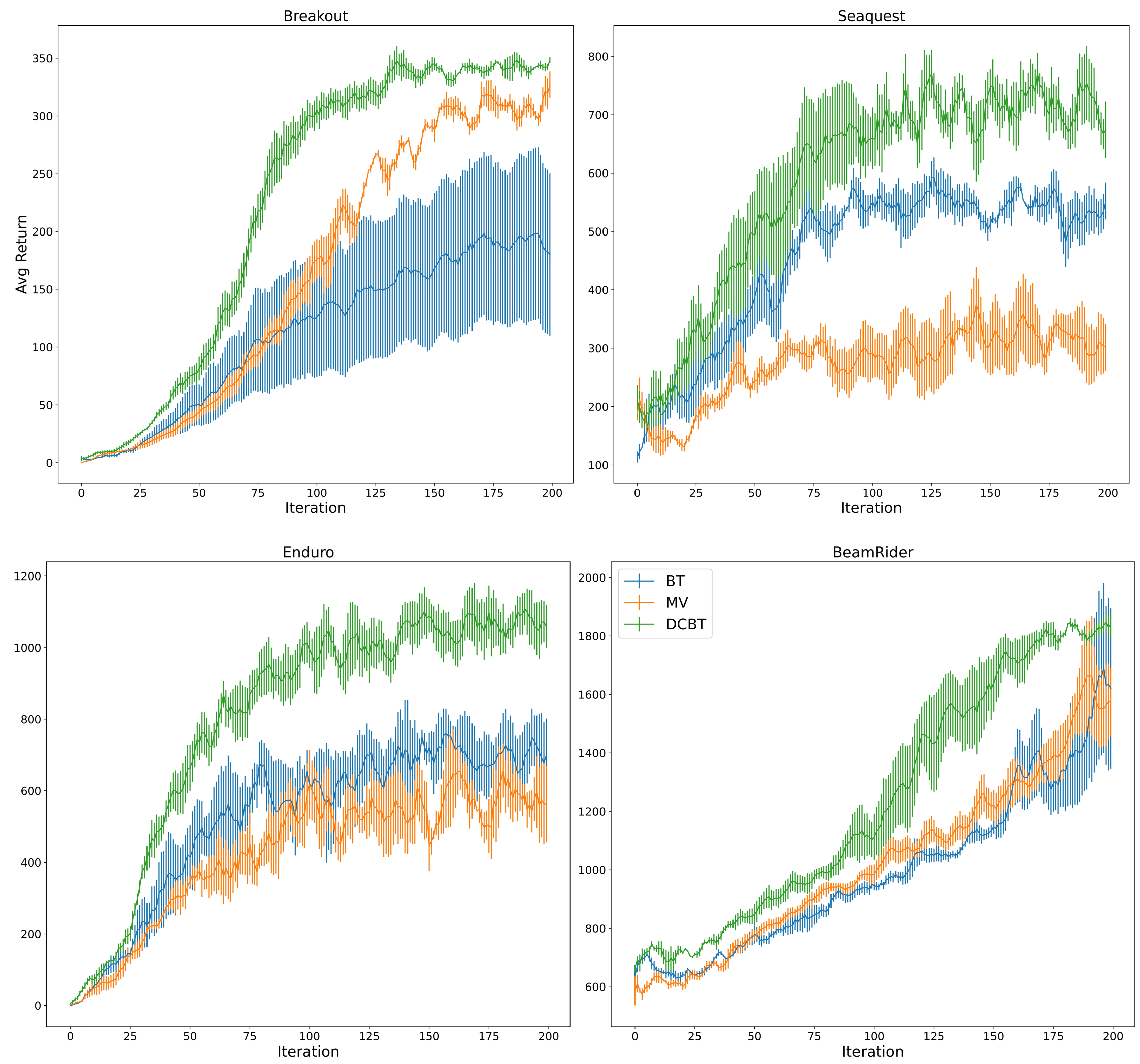
Experiment

- Generate synthetic annotators with sampled parameters for the probability of reporting correct answer.
- The proposed model is compared with the Bradley-Terry (BT) model used by previous work on preference-based RL and majority voting (MV).



For each dataset, learn rewards from the generated preferences, and then learn policies using the quantile-regression DQN algorithm.

The quality of learned policies reflects the performance of reward learning algorithms.



Conclusion

- MV cannot consistently outperforms BT, due to the fact that only a small amount of labels can be collected for each preference query.
- DCBT outperforms MV, which justifies using estimated P_{BT} and ID of annotators for denoising.
- DCBT achieves consistently good performance on all the four datasets, which confirms its efficacy and applicability.

Acknowledgement

This work was partially supported by JST CREST (JPMJCR21D1) and the RIKEN Guardian Robot Project.

Contact

Guoxi Zhang guoxi@ml.ist.i.kyoto-u.ac.jp

Hisashi Kashima kashima@i.kyoto-u.ac.jp