Internally Rewarded Reinforcement Learning

Mengdi Li*, Xufeng Zhao*, Jae Hee Lee, Cornelius Weber, Stefan Wermter

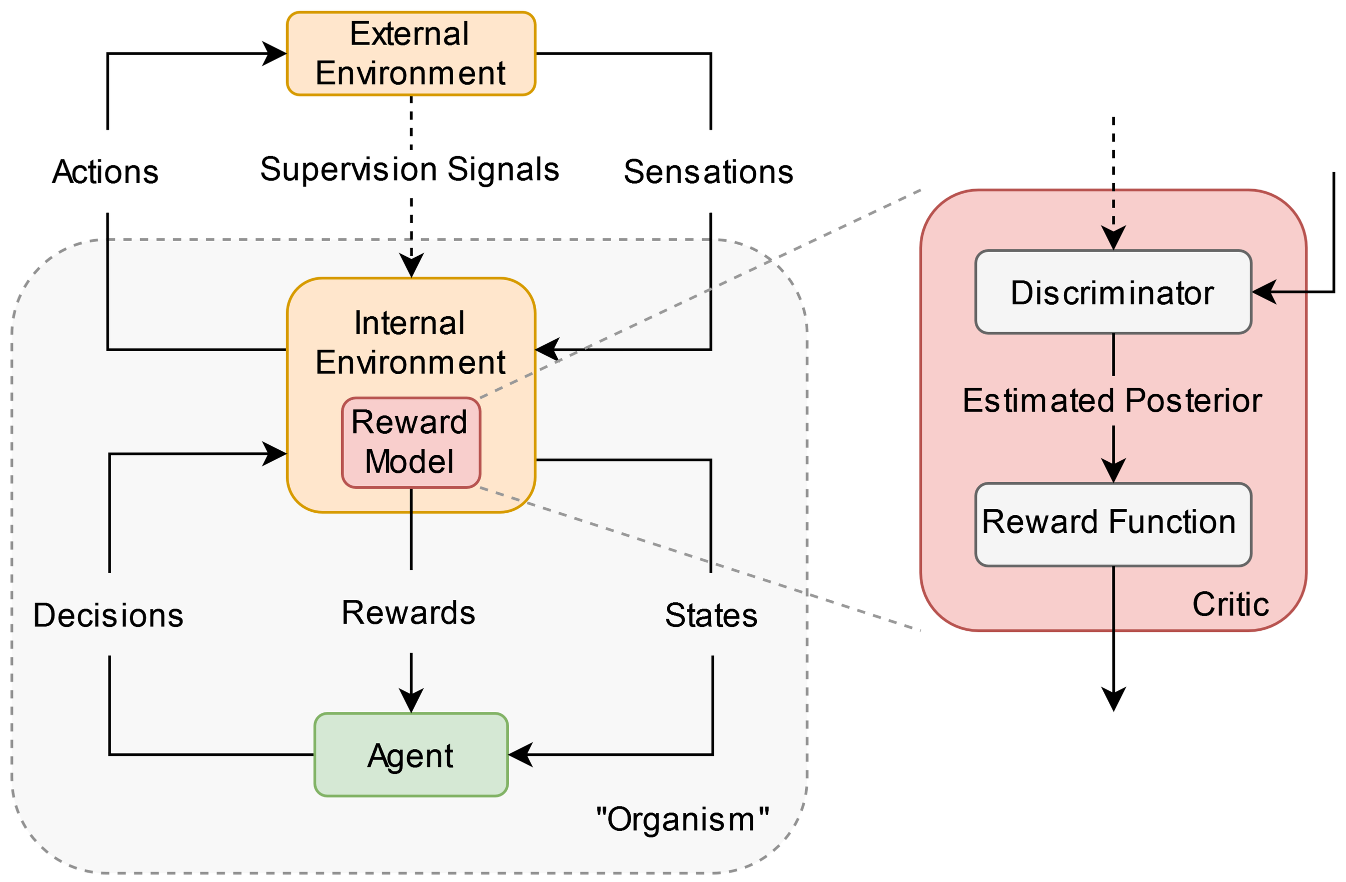
Knowledge Technology Group, University of Hamburg

www.knowledge-technology.info

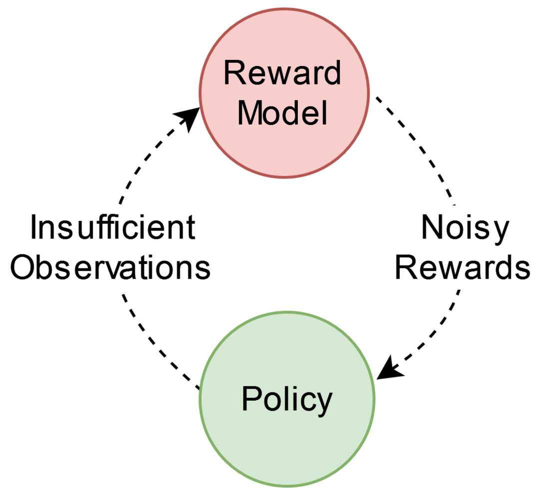
Summary

- We formulate ***internally rewarded RL*** (IRRL), where the reward for policy learning is *internally* provided by a discriminator that is dependent on and jointly optimized with the policy.
- We present the inherent ***issue of noisy rewards*** that results in an unstable training loop in IRRL.
- We propose a simple and effective reward function, ***the clipped linear reward function***, to reduce the impact of reward noise and stabilize the learning process.

Internally Rewarded RL



The agent-environment interaction loop of IRRL. Different from conventional RL settings, where rewards depend exclusively on the ***external*** environment, in IRRL rewards are determined by a ***reward model***, which resides in the ***internal*** environments.



- Simultaneous optimization between the ***policy*** of the agent and the ***reward model*** of the internal environment is challenging.
- An under-optimized reward model yields ***noisy rewards***, and in turn, an immature policy yields ***insufficient observations***, which leads to ***an unstable training loop***.

In IRRL, the policy and the discriminator are optimized simultaneously with ***different optimization objectives***:

Policy Optimization

Accuracy maximization:

$$r_{\text{acc}} = \mathbb{1}_y \left[\underset{y' \in \mathcal{Y}}{\operatorname{argmax}}\, q_\phi(y' \mid \tau) \right]$$

Mutual information maximization:

$$r_{\log} = \log q_\phi(y \mid \tau) - \log p(y)$$

Discriminator Optimization

Proxy cross-entropy loss:

$$-\mathbb{E}_{\tau \sim \pi_\theta, y \sim p(y)} \log q_\phi(y \mid \tau)$$

Notations

τ - Trajectory y - Label

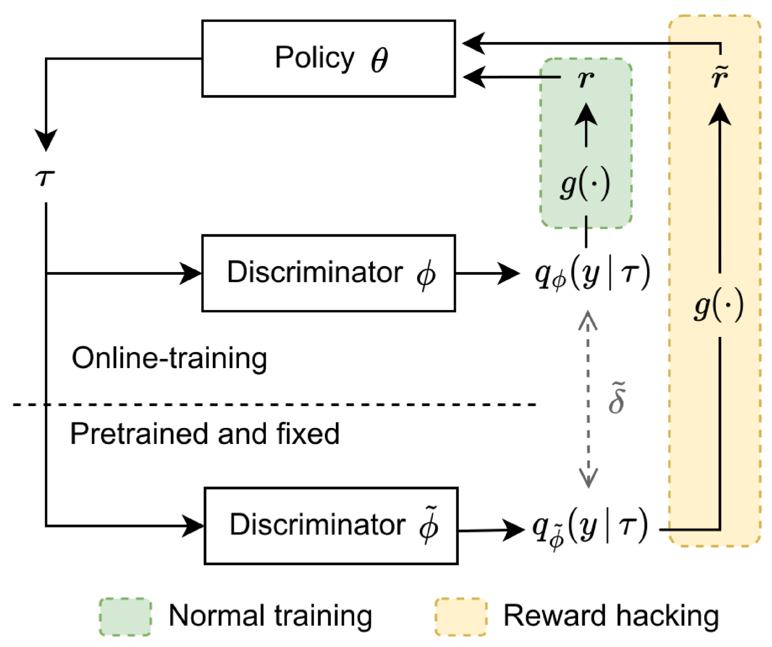
q_ϕ - Discriminator parameterized with ϕ

π_θ - Policy parameterized with θ

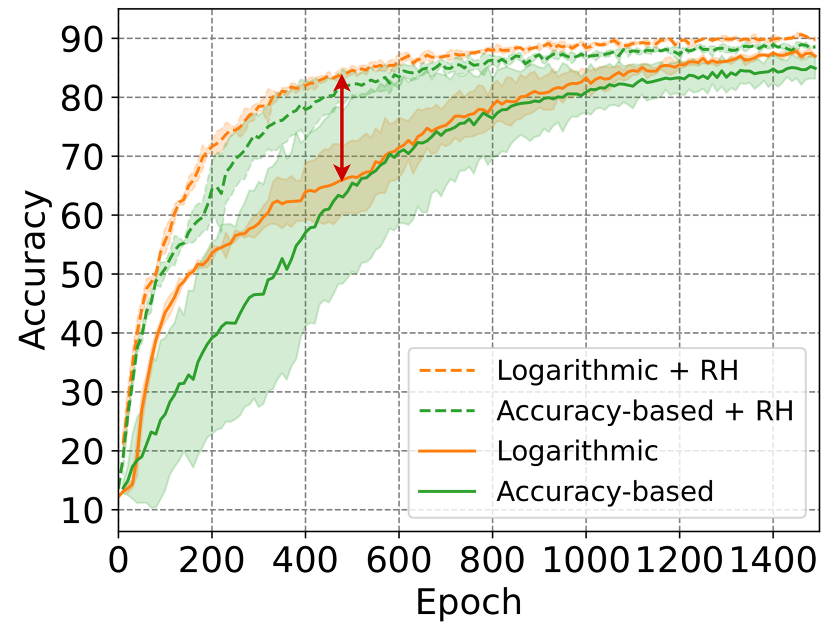
The Issue of Noisy Rewards

Oracle reward: $r^\star_{\log} = \log p(y \mid \tau) - \log p(y)$

Reward noise: $\varepsilon_{\log} = r_{\log} - r^\star_{\log} = \log q_\phi(y \mid \tau) - \log p(y \mid \tau)$



- Reward hacking: replace the ***trainable discriminator*** with a ***pretrained one*** that is from a converged model to mimic the ***oracle discriminator***.
- To demonstrate the impact of reward noise on the learning process.



- RAM trained using the accuracy-based and the logarithmic reward with and without ***reward hacking (RH)***.
- The ***gap*** between the training curves ***with and without reward hacking*** indicates the negative influence of noise from an under-optimized discriminator on the learning process.
- We aim to ***narrow the gap*** by moderating the reward noise.

Reward Noise Moderation

Since the noisy reward is a ***transformation*** of $q_\phi(y \mid \tau)$, it is reasonable to study the effect of the transformation as long as it results in the same optimal objective.

Generalized reward: $r_g = g\left[q_\phi(y \mid \tau)\right] - g\left[p(y)\right]$

Generalized reward noise: $\varepsilon_g := r_g - r^\star_g = g\left[q_\phi(y \mid \tau)\right] - g\left[p(y \mid \tau)\right]$

Discriminator noise: $\delta := q_\phi(y \mid \tau) - p(y \mid \tau)$

$$\mathbb{E}[\varepsilon_g] \approx g'(p(y \mid \tau))\mathbb{E}[\delta] + \frac{1}{2!}g''(p(y \mid \tau))\mathbb{E}\left[\delta^2\right]$$

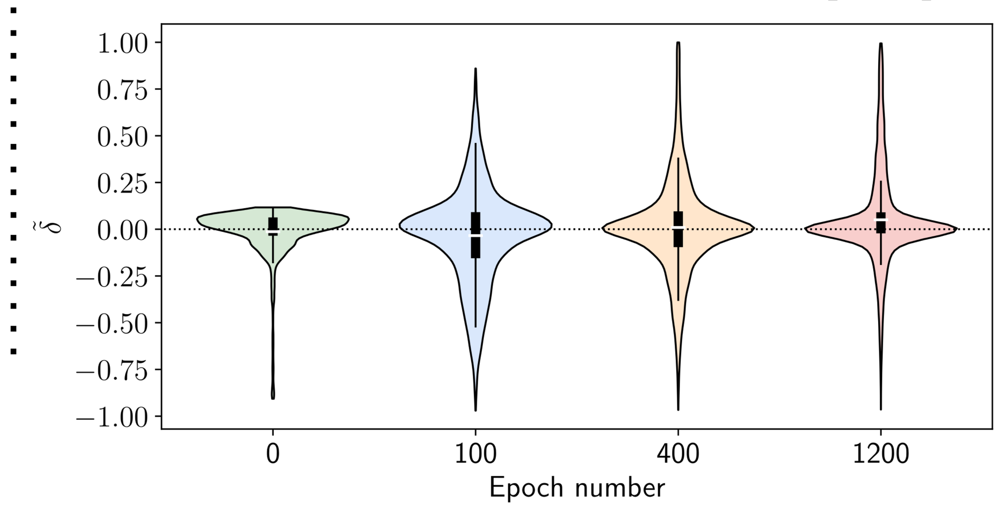
$$\mathbb{V}[\varepsilon_g] \approx (g'(p(y \mid \tau)))^2\mathbb{V}[\delta] + (\frac{1}{2!}g''(p(y \mid \tau)))^2\mathbb{V}\left[\delta^2\right]$$
$$+ g'(p(y \mid \tau))g''(p(y \mid \tau))\mathrm{Cov}\left[\delta, \delta^2\right]$$

Lower reward bias and ***more stable variance*** than the logarithmic reward

$$r_{\text{lin}} = q_\phi(y \mid \tau) - p(y)$$

$p(y \mid \tau)$ should be ***equal or larger*** than the prior $p(y)$.

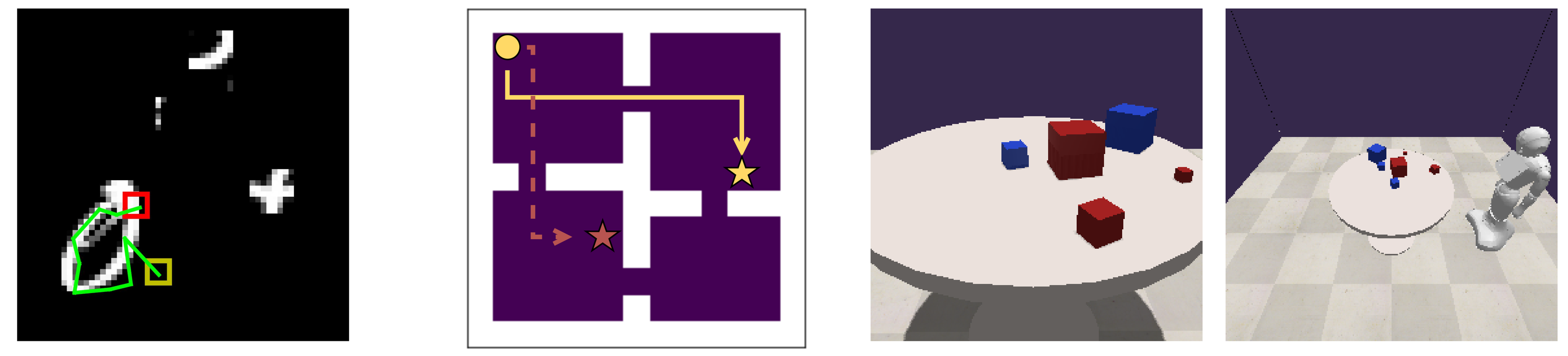
$$\overline{r_{\text{lin}}} = \max(q_\phi(y \mid \tau) - p(y), 0)$$



Characterization of the discriminator noise

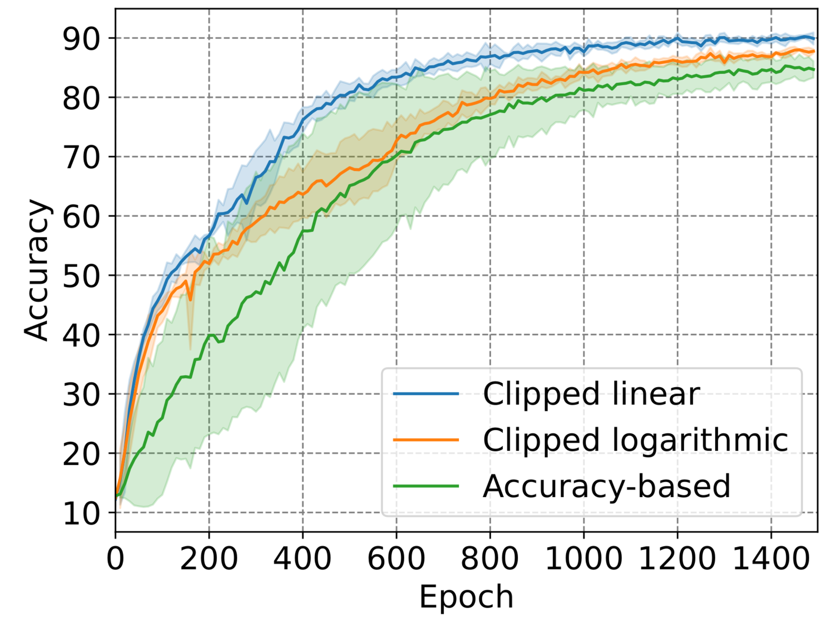
Experiments

We conduct experiments on three tasks:

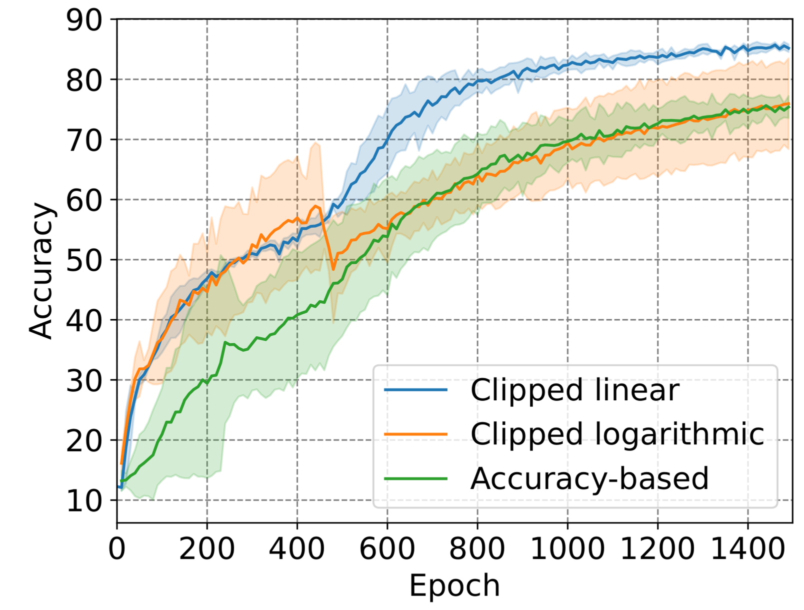


Hard attention for digit recognition Unsupervised Skill discovery Robotic object counting

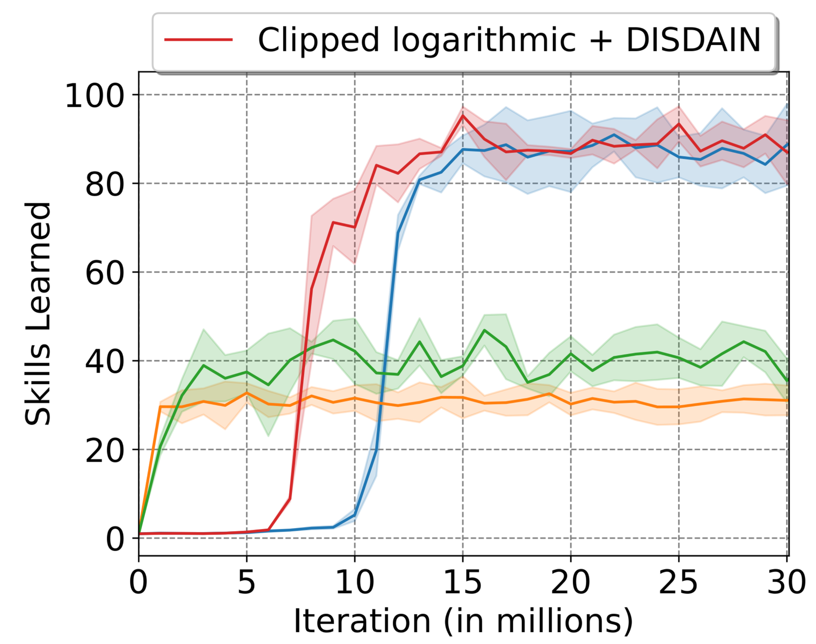
We compare the proposed ***clipped linear reward*** with alternative reward functions, including the ***clipped logarithmic reward*** and ***accuracy-based reward***. On the unsupervised skill discovery task, we additionally compare with the DISDAIN reward function.



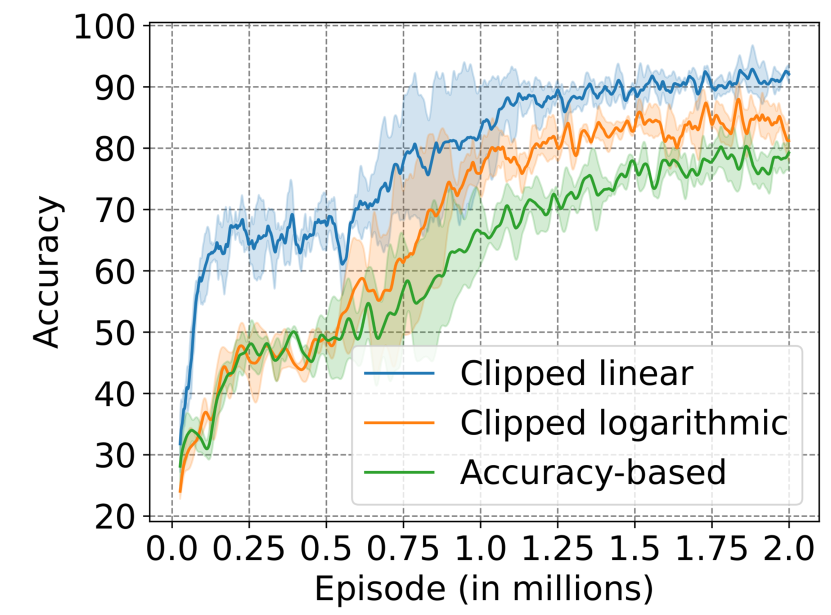
RAM on hard attention for digit recognition



DT-RAM on hard attention for digit recognition



Unsupervised skill discovery



Robotic object counting

The proposed clipped linear reward function consistently ***stabilizes the learning process*** and achieves faster convergence and higher performance compared with baselines in diverse tasks.

Reference

- Singh, S., Barto, A. G., and Chentanez, N. Intrinsically motivated reinforcement learning. In Advances in Neural Information Processing Systems (NeurIPS), 2004.
- Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. Recurrent models of visual attention. In Advances in Neural Information Processing Systems (NeurIPS), 2014.
- Strouse, D., Baumli, K., Warde-Farley, D., Mnih, V., and Hansen, S. Learning more skills through optimistic exploration. In International Conference on Learning Representations (ICLR), 2022.

ir-rl.github.io