Decoupling Exploration and Exploitation in Reinforcement Learning

Lukas Schäfer, Filippos Christianos, Josiah Hanna, Stefano V. Albrecht

Contact: l.schaefer@ed.ac.uk; Twitter: @LukasSchaefer96

ArXiv Code

Summary

- **Problem:** Intrinsic rewards in RL suffer from instability and sensitivity to hyperparameters
- **Idea:** Decouple exploration for data collection and training of an effective policy for exploitation.
- **Contributions:**
  1. Formulate on-policy and off-policy Decoupled RL (DeRL)
  2. Evaluate DeRL in sparse-reward environments with improved sample efficiency in several tasks
  3. Verify sensitivity of intrinsically motivated RL to scale and speed of decay of intrinsic rewards and demonstrate improved robustness of DeRL in two environments.

Motivation

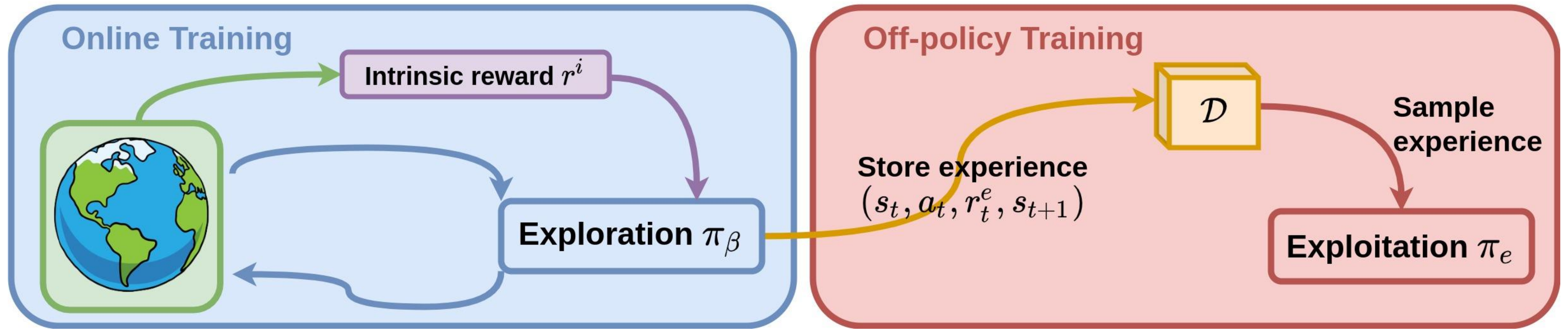
Intrinsic rewards: $r = r^e + \lambda r^i$

Intrinsic rewards are commonly applied to benefit exploration in RL. These approaches are particularly effective in environments where rewards of the environment are sparse. However, intrinsic rewards suffer from several key challenges.

Challenges of intrinsic rewards
1. Non-stationary reward shaping
2. Sensitive to scale of r^i
3. Sensitive to speed of decay of r^i

Balance of extrinsic (r^e) and intrinsic rewards (r^i) is needed!

Decoupled Reinforcement Learning (DeRL)



Exploration policy π_β trained online in environment

$$\pi_\beta \in argmax_\pi \, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(r_t^e + \lambda r_t^i\right) | a_t \sim \pi(s_t)\right]$$
$$= argmax_\pi \, \mathbb{E}\left[G_t^{e+i} \, | a_t \sim \pi(s_t)\right]$$

Exploitation policy π_e trained offline from $\mathcal{D}$

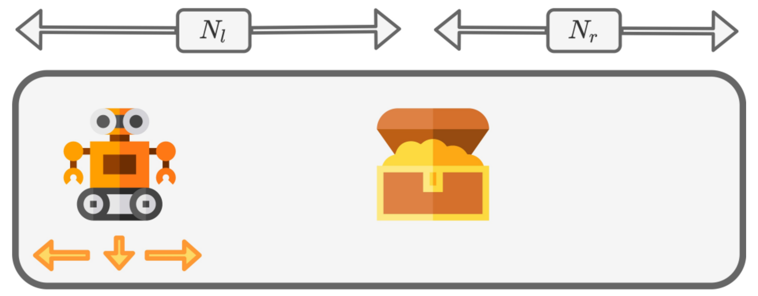
$$\pi_e \in argmax_\pi \, \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, r_t^e \, | (s_t, a_t, r_t^e, s_{t+1}) \sim \mathcal{D}\right]$$
$$= argmax_\pi \, \mathbb{E}\left[G_t^e \, || (s_t, a_t, r_t^e, s_{t+1}) \sim \mathcal{D}\right]$$

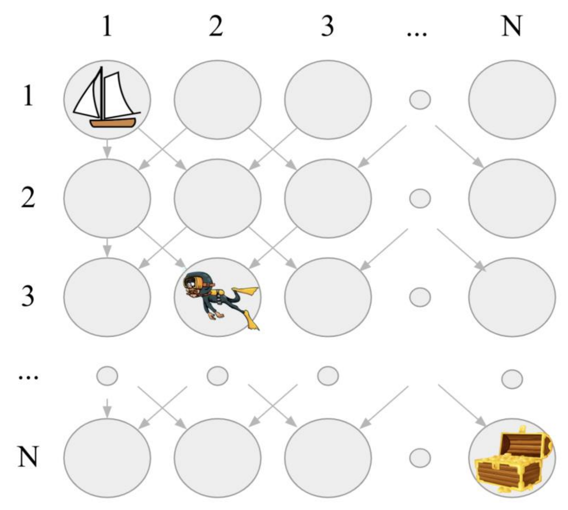
Decoupled On-Policy Learning

Optimise exploitation policy π_e using on-policy RL algorithms from experience samples $\mathcal{D}$. Needs off-policy correction like **importance sampling weights** $\rho(a_t|s_t) = \frac{\pi_e(a_t|s_t)}{\pi_\beta(a_t|s_t)}$.

Decoupled Off-Policy Learning

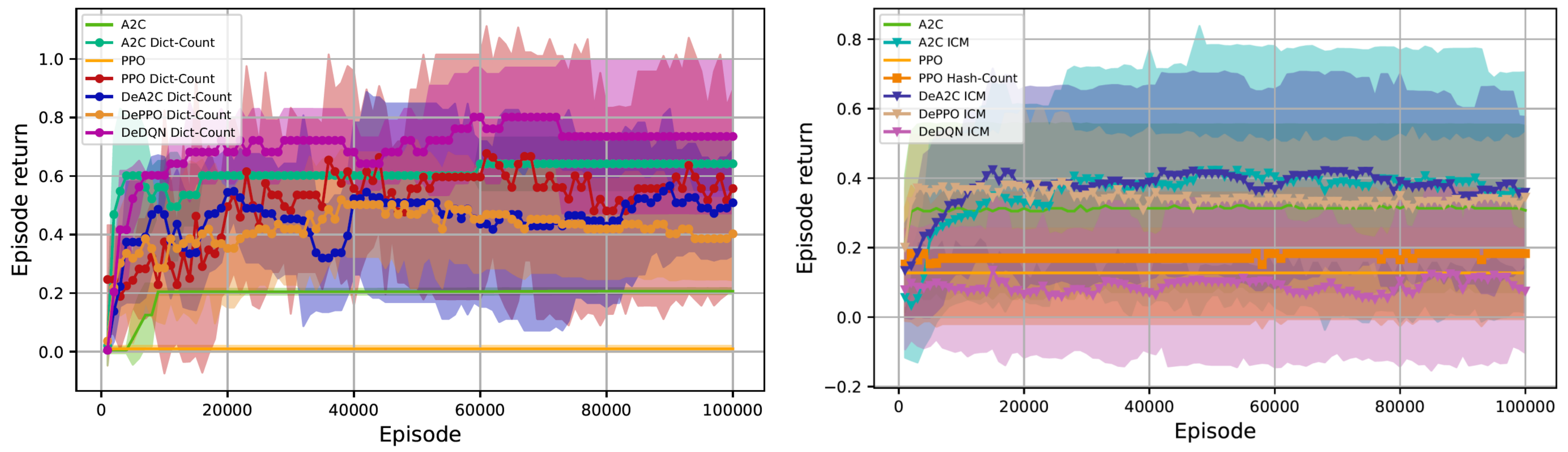
Optimise exploitation policy π_e using off-policy RL algorithms from experience samples $\mathcal{D}$. No off-policy correction needed and direct optimisation from $\mathcal{D}$ as a replay buffer.





Results

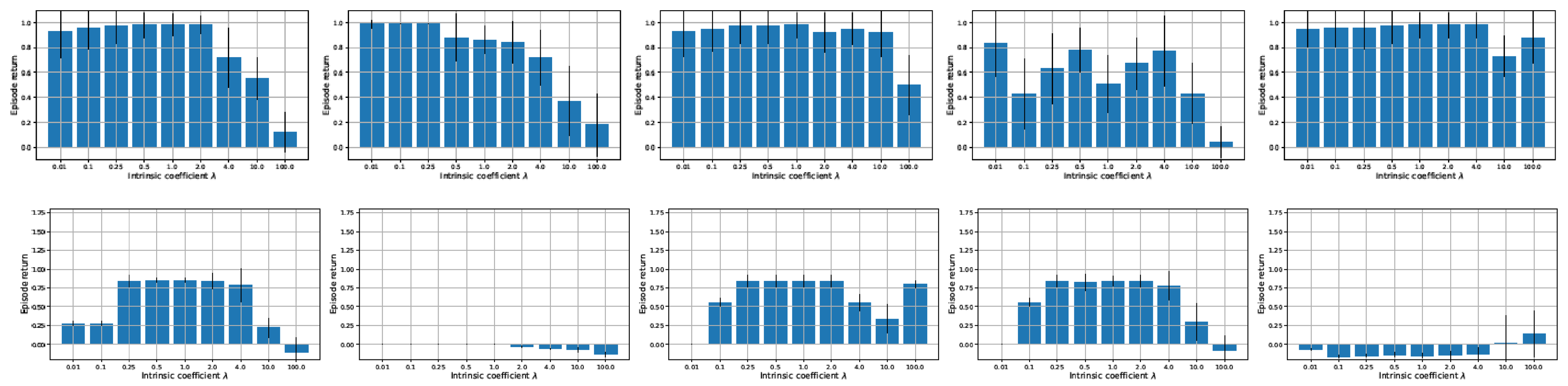
Normalised evaluation returns in DeepSea (left) and Hallway (right)



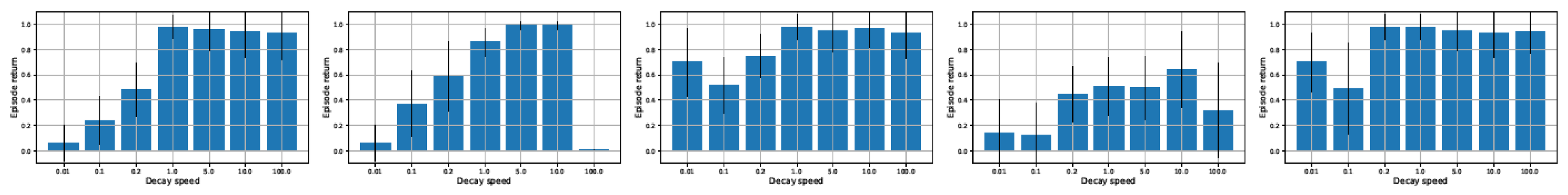
DeDQN and DeA2C outperform or match the best performing baselines in terms of converged returns and sample efficiency in the majority of DeepSea and simpler Hallway tasks.

Hyperparameter Sensitivity

Sensitivity of baselines and DeRL with Dict-Count intrinsic rewards to scale λ in



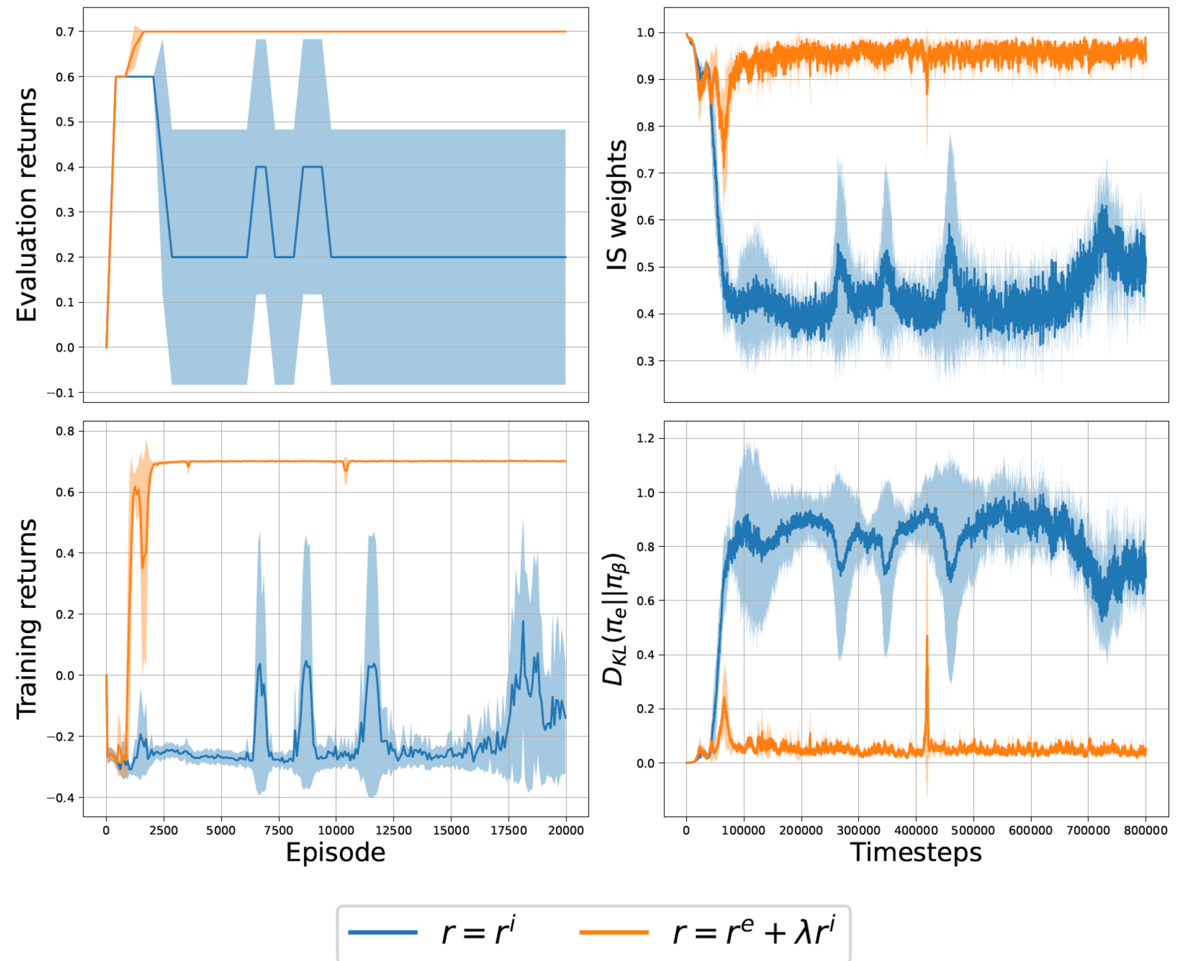
and speed of decay of intrinsic rewards in DeepSea $N = 10$



(a) A2C Dict-Count (b) PPO Dict-Count (c) DeA2C Dict-Count (d) DePPO Dict-Count (e) DeDQN Dict-Count

- **Scale and speed of decay of intrinsic rewards have a significant impact on returns.**
- **DeA2C and DeDQN are more robust to varying scale and speed of decay of intrinsic rewards in DeepSea and DeA2C in Hallway compared to baselines.**

Remaining Challenge: Distribution Shift



DeA2C optimised in Hallway ($N_l = N_r = 20$) with π_β trained using only intrinsic or combined rewards.

Exploration and exploitation policies diverge significantly throughout training (here DeA2C in Hallway 20-20). This effect is exacerbated if π_β is only trained with intrinsic rewards.

Future work:
Introduce divergence constraint to keep π_β and π_e close to each other:

$$D_{KL}\left(\pi_e \middle\| \pi_\beta\right)$$