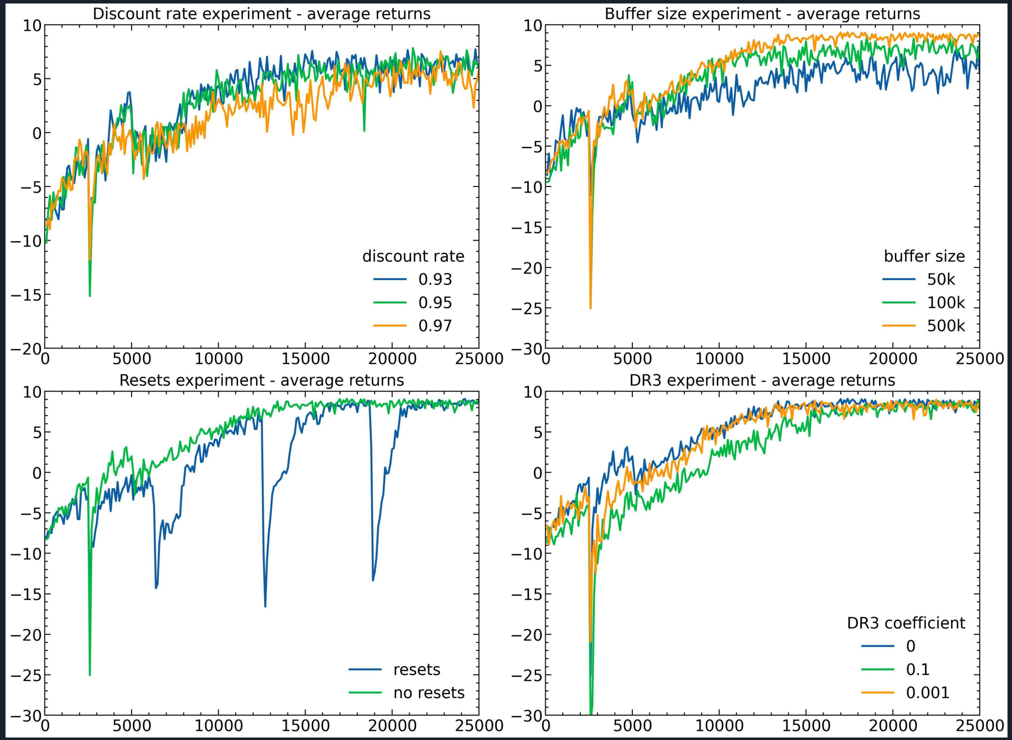
Reinforcement Learning Project SS23

RLcochet: Vera Milovanović, Khai Gandini, Filip Radović

Deep Deterministic Policy Gradient

Experiments:

- Discount rate tuning
- Buffer size tuning
- Network resetting
- DR3 regularization

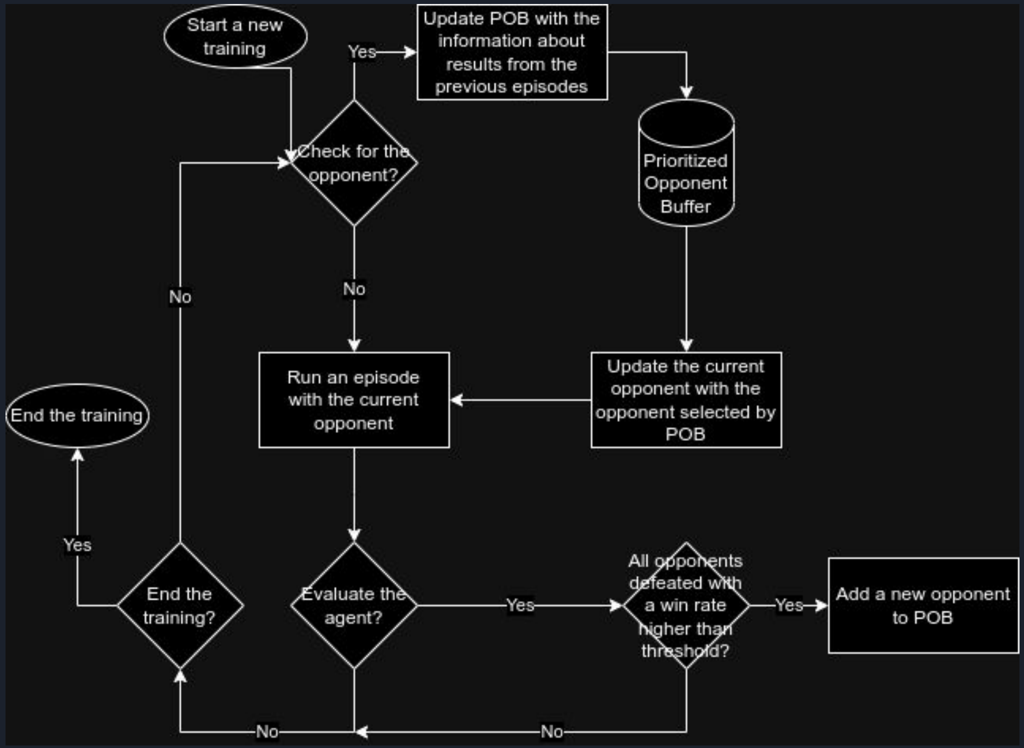


Deep Deterministic Policy Gradient - cont.

- Prioritized Opponent Buffer

Results:

- After 200 thousand episodes ended up with 8 opponents in the POB



TD3 (Twin-Delayed DDPG)

Improvements:

- Clipped Double Q-Learning (twin critics) $\left(Q(s,a) - (r + \gamma \min_{\boldsymbol{i=1,2}} \boldsymbol{Q_{\theta'_i}}(s', \pi(s')))\right)^2$
- Delayed Policy Updates
- Policy Smoothing Regularization $y = r + \gamma Q_{\text{target}}(s', \pi(s') + \epsilon)$

Extensions:

- Multi-Step Learning $R_t^{(n)} = \sum_{k=0}^{n-1} \gamma_t^{(k)} R_{t+k+1}$
- Prioritized Experience Replay (PER)

TD3 (Twin-Delayed DDPG)

Improvements:

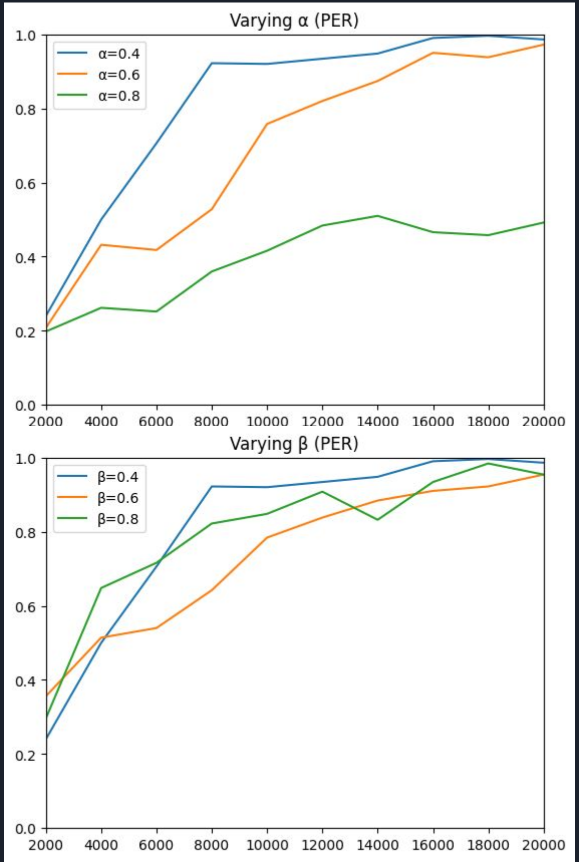
- Clipped Double Q-Learning (twin critics)
- Delayed Policy Updates
- Policy Smoothing Regularization

Extensions:

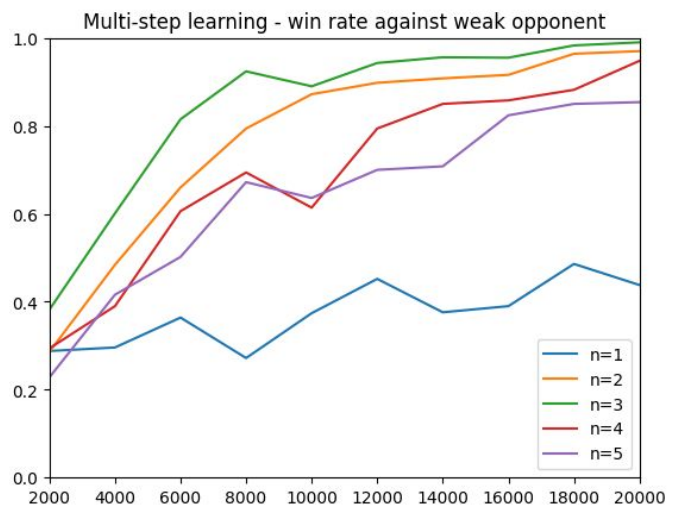
- Multi-Step Learning
- Prioritized Experience Replay (PER)

TD3 - Experiments

Prioritized Experience Replay



Multi-Step Learning (n-steps)



Soft Actor-Critic (SAC)

Main features:

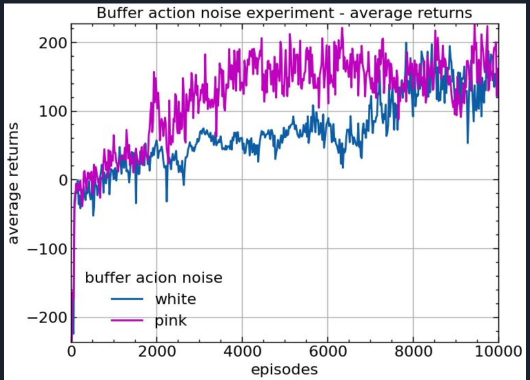
- Same architecture as TD3
- Off-policy algorithm that utilizes experience from the replay buffer
- Learning stochastic policy
- Objective function is regularized by the entropy of the policy => tackles exploration-exploitation problem (controlled by the temperature parameter)

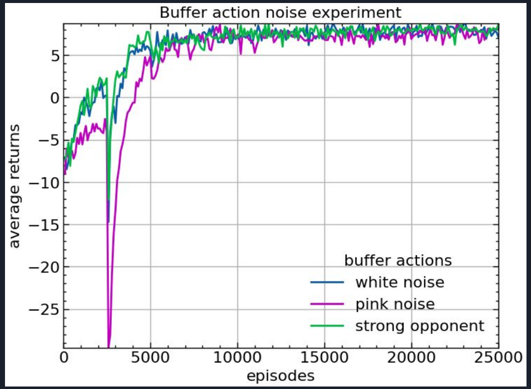
$$J(\theta) = \sum_{t=1}^{T} \mathbb{E}_{(s_t, a_t) \sim \rho_{\pi_\theta}} [r(s_t, a_t) + \alpha \mathcal{H}(\pi_\theta(.|s_t))]$$

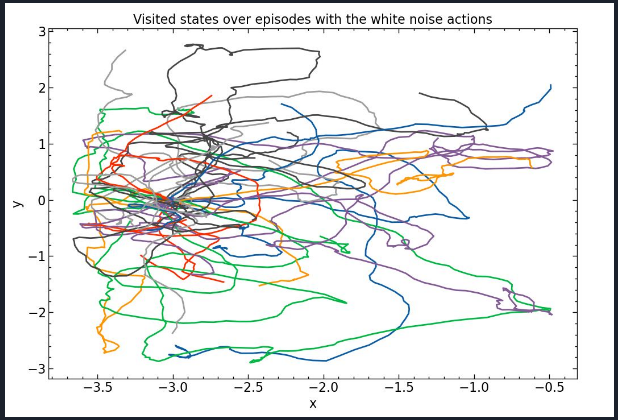
SAC extensions

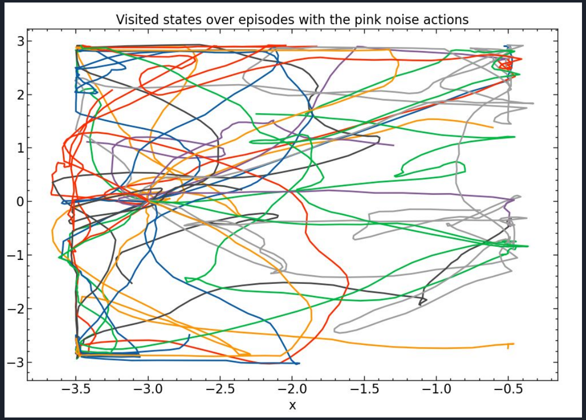
Pink noise

- Improve the early stage state space exploration by populating the replay buffer for the first 2000 episodes with the actions sampled from the pink process









SAC training process for Hockey environment

Three modes of playing: train defense, shooting and normal play

Training pipeline:

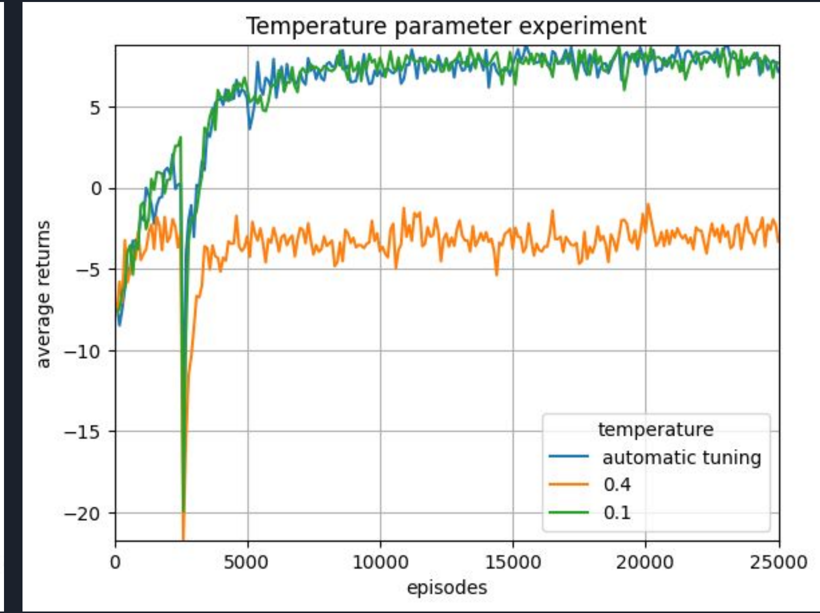
- Train defense for 2500 episodes
- Train shooting for next 2500 episodes
- For each episode choose any of the three modes randomly

In parallel use POB for sampling the opponents

Results:

- Improved robustness - agent doesn’t overfit the opponents
- After 175 thousand episodes ended up with 8 opponents in the POB

SAC - hyperparameter tuning
Two most influential hyperparameters



Agents in action

Thank you for listening!