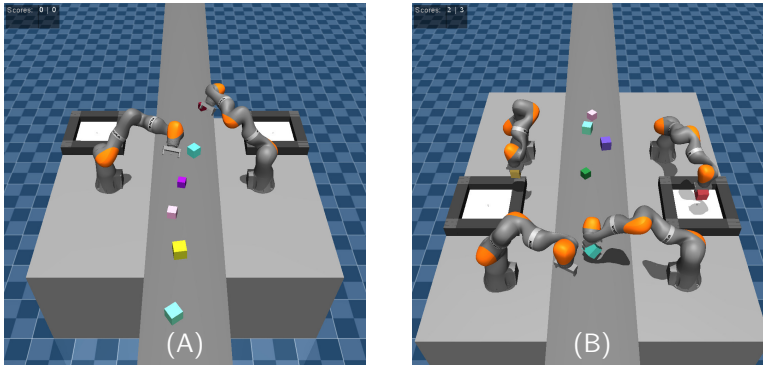


# Cooperative Multi-agent RL for Factory Manipulation

Nikolas Kirschstein & Kassian Köck (Team 7)



## Big Picture

**goal:** multiple robot arms *cooperating* to *maximise efficiency* in factory manipulation task like pick-and-place  
**issue:** conventional pre-programming too *inflexible* and *tedious*  
**conjecture:** *multi-agent RL* may find near-optimal behaviour

## Given: The Environment

- even number of robot arms (8 DOF each)
- basket in reach for each arm
- conveyor belt with increasing speed transporting cubes
- score = number of cubes in baskets
- episode termination if either:
  - arm hits the environment (incl. other arms)
  - cube is missed by all arms

## Given: IK Base Policy

- state machine working on *one target object at a time*
- different base policies *ignore each others'* target objects
- control calculated via *inverse kinematics (IK)*

## (A) Continuous Control: Setting

**action space:**  $[-1, 1]$  per joint and learnt arm to control joint state  
**observation space:** joint and cube states

learning choices per arm:

- *full RL*: learn entire joint control RL from scratch (hard)
- *delta*: learn only deviation from IK base policy (simplification)
- *base*: execute IK base policy (baseline)

## (A) Continuous Control: Reward Shaping

- problem: learning hard due to discrete, *highly sparse* reward  
⇒ denser reward function needed
- desirable incentives:
  - $I_0$ : reward *new cubes* put into bucket
  - $I_1$ : reward approach of *gripper* to closest cube
  - $I_2$ : reward approach of closest cube to *bucket*
- reward function:  $r = r_0 + \omega_0 I_0 + \omega_1 I_1 + \omega_2 I_2$   
(base reward  $r_0$  prevents learning to terminate episode)

## (A) Continuous Control: Results

for 2 robot arms (simplest case):

- pure sparse reward: absolutely no learning, *random behaviour*
- progress-based reward ( $r_0 = 0.4, \omega_0 = 1, \omega_1 = 0.2, \omega_2 = 0.4$ )
  - successful *collision avoidance* due to implicit survival reward
  - BUT: *no gripping* at all (see screenshot (A) and videos)

test results averaged over 100 episodes

learning choice	avg episode score	avg episode length
<i>full RL</i>	(0, 0)	219.9
<i>delta</i>	(0, 0)	220.5
<i>base</i>	(1.65, 1.17)	208.8

## (B) Discrete Control: Setting

**action space:** {ON, OFF} per arm to toggle use of IK base policy  
**observation space:** joint and cube states + *proposed IK control*

behaviour choices for case OFF:

- *pause*: freeze at the current position
- *retreat*: return to a safe default position
- *base*: continue executing IK policy

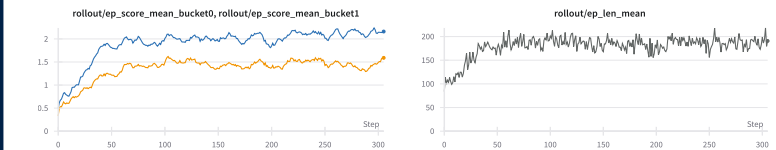
## (B) Discrete Control: Results

for case of 4 robot arms:

- *pure sparse score-based reward* suffices for effective learning!
- successful *dodging and gripping*! (see screenshot (B) and videos)
- pausing strategy manages to *improve upon baseline*!

test results averaged over 100 episodes

learning choice	avg episode score	avg episode length
<i>pause</i>	(2.01, 1.5)	173.04
<i>retreat</i>	(0, 2.12)	216.74
<i>base</i>	(1.18, 1.16)	119.74



## Future Work

- modify setting s.t. discrete control decides *which cube to grab*
- introduce *auxiliary tasks* (grip, carry, release) & *curriculum learning*  
⇒ requires careful design and much more computing resources