Reward, observation and action shapes used in the training environments

1 Observation

OBS1

$$[Ex, Ey, Ez, A1, A2, A3, A4, A5, A6] \tag{1}$$

OBS2

$$[Gx, Gy, Gz, A1, A2, A3, A4, A5, A6] \tag{2}$$

OBS3

$$[ETx, ETy, ETz, EGx, EGy, EGz, A1, A2, A3, A4, A5, A6] \tag{3}$$

OBS4

$$[EGx, EGy, EGz, A1, A2, A3, A4, A5, A6] \tag{4}$$

OBS5

$$[ETx, ETy, ETz, EGx, EGy, EGz, Gx, Gy, Gz, A1, A2, A3, A4, A5, A6] \tag{5}$$

where

- Ei : End effector coordinate along the i axis
- Gi : Goal coordinate along the i axis
- EGi : Vector End effector - Goal along the i axis
- ETx : Vector End effector - Torso along the i axis
- Ai : Angular position of joint i

2 Reward

2.1 Dense reward functions

REW1

$$r = -d_t^2 \tag{6}$$

REW2

$$r = -d_t \tag{7}$$

REW3

$$r = -d_t^3 \tag{8}$$

REW4

$$r = -d_t^4 \tag{9}$$

REW5

$$r = -d_t^2 - \alpha \left\| A_t \right\| \tag{10}$$

REW6

$$r = -d_t^2 - \alpha \frac{\|A_t\|}{d_t^2} \tag{11}$$

REW7

$$r = \Delta d_t \tag{12}$$

REW8

$$r = -d_t^2 + \alpha abs(\frac{\Delta d_t}{d_t^2}) \tag{13}$$

REW9

$$r = \Delta E_t \tag{14}$$

REW10

$$r = -d_t^2 + \alpha \frac{\Delta E_t}{d_t^2} \tag{15}$$

2.2 Sparse reward functions

REW11

$$r = \begin{cases} -1, & \text{if } d_t \geq \epsilon \\ 0, & \text{if } d_t < \epsilon \end{cases} \tag{16}$$

REW12

$$r = \begin{cases} 1, & \text{if } d_t \geq \epsilon \\ 0, & \text{if } d_t < \epsilon \end{cases} \tag{17}$$

REW13

$$r = \begin{cases} -0.02, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{18}$$

REW14

$$r = \begin{cases} -0.001, & \text{if } d \geq \epsilon \\ 10, & \text{if } d < \epsilon \end{cases} \tag{19}$$

2.3 Sparse + dense reward functions

REW15: BEST REWARD FUNCTION FOR DISTANCE

$$r = \begin{cases} -d_t, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{20}$$

REW16

$$r = \begin{cases} \Delta d_t, & \text{if } d \geq \epsilon \\ \Delta d_t + 10, & \text{if } d < \epsilon \end{cases} \tag{21}$$

2.4 Position + orientation: Dense reward functions

REW17

$$r = -O_t^2 \tag{22}$$

REW18

$$r = -d_t^2 - O_t^2 \tag{23}$$

REW19

$$r = \begin{cases} -d_t^2 - \alpha O_t^2, & \text{if } d_t \geq 0.001 \text{ and } O_t \geq 0.01 \\ 1, & \text{if } d_t < 0.001 \text{ and } O_t < 0.01 \end{cases} \tag{24}$$

2.5 Sparse + dense reward functions + Penalty for collision

REW20

$$r = \begin{cases} -d_t, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{25}$$

+ Penalty -10 if collision detected.
where

- r : Reward
- d_t : Distance at time t
- O_t : Orientation vector (collinearity between the end effector and the goal orientation)
- Δd_t : Change in distance
- a_t : Action at time t
- A_t : Action normalised between -1 and 1
- E_t : End effector position at time t
- ΔE_t : Change in position
- α : Scaling coefficient (0.1)
- ϵ : Threshold for sparse reward (0.001)

2.6 Dense rewards (from the literature)

$$r = -d_t^2 \tag{26}$$

$$r = -d_t \tag{27}$$

$$r = -\alpha d_t - \beta a^T a \tag{28}$$

$$r = -\alpha d_{t-1}^p - d_t^p \tag{29}$$

$\alpha = 0$ or 1
$p = 1$ or 2
but don't work well...

$$r = -d_t - \|a_{t-1}\| \tag{30}$$

Penalise large torque

$$r = -d_t^2 + \frac{d_{t-1} - d_t}{d_t} \tag{31}$$

2.7 Sparse rewards (from the literature)

$$r = \begin{cases} -1, & \text{if } d \geq \epsilon \\ 0, & \text{if } d < \epsilon \end{cases} \tag{32}$$

$$r = \begin{cases} 1, & \text{if } s \in G \\ 0, & \text{otherwise} \end{cases} \tag{33}$$

2.8 Dense + sparse rewards (from the literature)

$$r = \begin{cases} -d_t, & \text{if no collision and } d \geq 3 \\ -d_t - 20\beta, & \text{if collision and } d \geq 3 \\ -d_t + 2, & \text{if no collision and } d < 3 \\ -d_t - 20\beta + 2, & \text{if collision and } d < 3 \end{cases} \tag{34}$$

$$r = \begin{cases} -1 - \beta \left\| a_{t-1} \right\|^2, & \text{if } d \geq \epsilon \\ 1 - \beta \left\| a_{t-1} \right\|^2, & \text{if } d < \epsilon \end{cases} \tag{35}$$

where $\beta \left\| a_{t-1} \right\|^2 \ll 1$ (penalise large actions)

$$r = \begin{cases} -d_t, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{36}$$

$$r = \begin{cases} -0.02, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{37}$$

$$r = \begin{cases} \alpha(d_{t-1} - d_t), & \text{if } d \geq \epsilon \\ \alpha(d_{t-1} - d_t) + 10, & \text{if } d < \epsilon \end{cases} \tag{38}$$

$$r = \begin{cases} -0.001, & \text{if } d \geq \epsilon \\ 10, & \text{if } d < \epsilon \end{cases} \tag{39}$$

Where s = state
G = set of goals

3 Action

ACT1 : Relative joint position (Instant reset of joint positions)

$$[\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6] \tag{40}$$

ACT2 : Relative joint position (Continuous position control and compute physics for collision)

$$[\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6] \tag{41}$$

ACT3 : Absolute joint position
ACT4 : Relative joint torque
ACT5 : Absolute joint torque
Where δ_i : Increment from previous joint position (in rad)