

Reward, observation and action shapes used in the training environments

1 Observation

OBS1

$$[Ex, Ey, Ez, A1, A2, A3, A4, A5, A6] \quad (1)$$

OBS2

$$[Gx, Gy, Gz, A1, A2, A3, A4, A5, A6] \quad (2)$$

OBS3

$$[ETx, ETy, ETz, EGx, EGY, EGz, A1, A2, A3, A4, A5, A6] \quad (3)$$

OBS4

$$[EGx, EGY, EGz, A1, A2, A3, A4, A5, A6] \quad (4)$$

OBS5

$$[ETx, ETy, ETz, EGx, EGY, EGz, Gx, Gy, Gz, A1, A2, A3, A4, A5, A6] \quad (5)$$

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

$$r = -d_t^2 \quad (6)$$

REW2

$$r = -d_t \quad (7)$$

REW3

$$r = -d_t^3 \quad (8)$$

REW4

$$r = -d_t^4 \quad (9)$$

REW5

$$r = -d_t^2 - \alpha \|A_t\| \quad (10)$$

REW6

$$r = -d_t^2 - \alpha \frac{\|A_t\|}{d_t^2} \quad (11)$$

REW7

$$r = \Delta d_t \quad (12)$$

REW8

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

REW9

$$r = \Delta E_t \quad (14)$$

REW10

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

2.2 Sparse reward functions

REW11

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

REW12

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

REW13

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

REW14

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

2.3 Sparse + dense reward functions

REW15: BEST REWARD FUNCTION FOR DISTANCE

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

REW16

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

2.4 Position + orientation: Dense reward functions

REW17

$$r = -O_t^2 \quad (22)$$

REW18

$$r = -d_t^2 - O_t^2 \quad (23)$$

REW19

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

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)

2.5 Dense rewards (from the literature)

$$r = -d_t^2 \quad (25)$$

$$r = -d_t \quad (26)$$

$$r = -\alpha d_t - \beta a^T a \quad (27)$$

$$r = -\alpha d_{t-1}^p - d_t^p \quad (28)$$

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

$$r = -d_t - \|a_{t-1}\| \quad (29)$$

Penalise large torque

$$r = -d_t^2 + \frac{d_{t-1} - d_t}{d_t} \quad (30)$$

2.6 Sparse rewards (from the literature)

$$r = \begin{cases} -1, & \text{if } d \geq \epsilon \\ 0, & \text{if } d < \epsilon \end{cases} \quad (31)$$

$$r = \begin{cases} 1, & \text{if } s \in G \\ 0, & \text{otherwise} \end{cases} \quad (32)$$

2.7 Dense + sparse rewards (from the literature)

$$r = \begin{cases} -d_t, & \text{if no collision and } d \geq 3 \\ -d_t - 20\beta, & \text{if collision and } d \geq 3 \\ -d_t + 2, & \text{if no collision and } d < 3 \\ -d_t - 20\beta + 2, & \text{if collision and } d < 3 \end{cases} \quad (33)$$

$$r = \begin{cases} -1 - \beta \|a_{t-1}\|^2, & \text{if } d \geq \epsilon \\ 1 - \beta \|a_{t-1}\|^2, & \text{if } d < \epsilon \end{cases} \quad (34)$$

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

$$r = \begin{cases} -d_t, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \quad (35)$$

$$r = \begin{cases} -0.02, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \quad (36)$$

$$r = \begin{cases} \alpha(d_{t-1} - d_t), & \text{if } d \geq \epsilon \\ \alpha(d_{t-1} - d_t) + 10, & \text{if } d < \epsilon \end{cases} \quad (37)$$

$$r = \begin{cases} -0.001, & \text{if } d \geq \epsilon \\ 10, & \text{if } d < \epsilon \end{cases} \quad (38)$$

Where s = state

G = set of goals

3 Action

ACT1 : Relative joint position

$$[\delta_1, \delta_2, \delta_3, \delta_4, \delta_5, \delta_6] \quad (39)$$

ACT2 : Absolute joint position

ACT3 : Relative joint torque

ACT4 : Absolute joint torque

Where δ_i : Increment from previous joint position (in rad)

TODO: Also make the difference between immediate reset and continuous position control.