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

REW1

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

REW2

$$r = -d_t^2 - \alpha \left\| a_t \right\| \tag{7}$$

REW3

$$r = d_{t-1} - d_t \tag{8}$$

REW4

$$r = -d_t^2 - \alpha \frac{d_{t-1} - d_t}{d_t} \tag{9}$$

REW5

$$r = \begin{cases} -1, & \text{if } d \geq \epsilon \\ 0, & \text{if } d < \epsilon \end{cases} \tag{10}$$

REW6

$$r = \begin{cases} 1, & \text{if } d \geq \epsilon \\ 0, & \text{if } d < \epsilon \end{cases} \tag{11}$$

REW7

$$r = 1 - abs(A_t[0]) \tag{12}$$

where

- r : Reward
- d_t : Distance at time t
- a_t : Action at time t
- A_t : Action normalised between -1 and 1
- α : Scaling coefficient (1)
- ϵ : Threshold for sparse reward (0.001)

2.1 From the literature

2.1.1 Dense rewards

$$r = -d_t^2 \tag{13}$$

$$r = -d_t \tag{14}$$

$$r = -\alpha d_t - \beta a^T a \tag{15}$$

$$r = -\alpha d_{t-1}^p - d_t^p \tag{16}$$

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

$$r = -d_t - \|a_{t-1}\| \tag{17}$$

Penalise large torque

$$r = -d_t^2 + \frac{d_{t-1} - d_t}{d_t} \tag{18}$$

2.1.2 Sparse rewards

$$r = \begin{cases} -1, & \text{if } d \geq \epsilon \\ 0, & \text{if } d < \epsilon \end{cases} \tag{19}$$

$$r = \begin{cases} 1, & \text{if } s \in G \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

2.1.3 Dense + sparse rewards

$$r = \begin{cases} -d_t, & \text{if no collision and } d \geq 3 \\ -d_t - 20\beta, & \text{if collision and } d \geq 3 \\ -d_t + 2, & \text{if no collision and } d < 3 \\ -d_t - 20\beta + 2, & \text{if collision and } d < 3 \end{cases} \tag{21}$$

$$r = \begin{cases} -1 - \beta \left\| a_{t-1} \right\|^2, & \text{if } d \geq \epsilon \\ 1 - \beta \left\| a_{t-1} \right\|^2, & \text{if } d < \epsilon \end{cases} \tag{22}$$

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

$$r = \begin{cases} -d_t, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{23}$$

$$r = \begin{cases} -0.02, & \text{if } d \geq \epsilon \\ 1, & \text{if } d < \epsilon \end{cases} \tag{24}$$

$$r = \begin{cases} \alpha(d_{t-1} - d_t), & \text{if } d \geq \epsilon \\ \alpha(d_{t-1} - d_t) + 10, & \text{if } d < \epsilon \end{cases} \tag{25}$$

$$r = \begin{cases} -0.001, & \text{if } d \geq \epsilon \\ 10, & \text{if } d < \epsilon \end{cases} \tag{26}$$

$$r = \begin{cases} -0.001, & \text{if } d \geq \epsilon \\ 10, & \text{if } d < \epsilon \end{cases} \tag{27}$$

Where s = state
G = set of goals

3 Action

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

ACT1 : Relative joint position

$$[\delta_1, delta_2, delta_3, delta_4, delta_5, delta_6] \tag{28}$$

ACT2 : Absolute joint position

ACT3 : Relative joint torque

ACT4 : Absolute joint torque