Learning from diverse feedback types

- demonstrations: ξ_D
- comparisons (preferences): ξ_A or ξ_B ?
- physical corrections $\tilde{C}$
- turning the robot off ⏻
- saying something $\lambda =$ "don't step on the carpet
- specifying a reward? $\tilde{\theta}$
- the current state of the world s_0

$$b'(\theta) \propto b(\theta) \underbrace{P(\text{feedback} \mid \theta)}_{\text{human model}}$$

choices $c \in C$

choose based on reward

$R_\theta(c)$? ×

$R_\theta(\psi(c))$

$\hat{}$ grounding of c into traj

$\psi : c \mapsto \xi$

OR $\psi : c \mapsto \Delta$ $\mathbb{E}[R_\theta(\xi) \mid \xi \sim \psi(c)]$

$$P(c^* \mid \theta) = \frac{e^{R_\theta(\psi(c^*))}}{\sum_{c \in C} e^{R_\theta(\psi(c))}}$$

demonstrations

$$C = \{\xi\} \quad \psi(\xi) = \xi \quad P(\xi_D \mid \theta) = \frac{e^{R_\theta(\xi_D)}}{\sum_\xi e^{R_\theta(\xi)}}$$

comparisons

$$C = \{\xi_A, \xi_B\} \quad \psi(\xi) = \xi \quad P(\xi_A \mid \theta) = \frac{e^{R_\theta(\xi_A)}}{e^{R_\theta(\xi_A)} + e^{R_\theta(\xi_B)}}$$

$$= \frac{1}{1 + e^{R_\theta(\xi_B) - R_\theta(\xi_A)}}$$

off switch

$$C = \{①, -\} \quad \psi(-) = \xi_R \quad \psi(①) = \xi_D^{0:t} \xi_R^{t} \dots \xi_R^{t}$$

corrections

$$C = \{\Delta q\} \quad \psi(\Delta q) = \xi_R + A^{-1} \Delta q$$

(proxy) specified reward

$$C = \{\theta\} \quad \psi(\theta) \sim P(\xi \mid \theta, \underline{M_{deuce}})$$

current state

$C = \{s\} \quad \psi(s) \sim \mathrm{Unif}(S_H^{-T:0} \mid S_H^0 = s)$