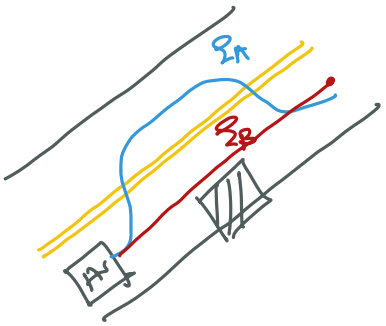
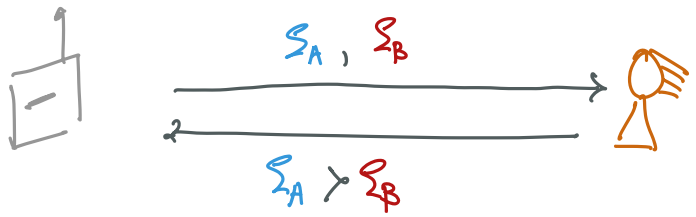


TRAJECTORY $\xi = \{ s_0, a_0, s_1, a_1, s_2, a_2, \ldots \}$



PREFERENCE DATASET

$$\left\{ \begin{array}{lll} \xi_A^1, & \xi_B^1, & \succ \\ \xi_A^2, & \xi_B^2, & \prec \\ \vdots & & \\ \xi_A^N, & \xi_B^N, & \succ \end{array} \right\}$$

Q: HOW CAN I LEARN A REWARD FUNCTION $R_\theta(s,a)$ FROM PREFERENCE?

BRADLEY TERRY MODEL

* PREFERENCES ARE PROBABILISTIC

POSITIVE, REAL-VALUED

$$P(A \succ B) = \frac{\text{SCORE}(A)}{\text{SCORE}(A) + \text{SCORE}(B)}$$

$$P(\xi_A \succ \xi_B) = \frac{\exp(R(\xi_A))}{\exp(R(\xi_A)) + \exp(R(\xi_B))}$$

$$= \frac{1}{1 + \exp(R(\xi_B) - R(\xi_A))}$$

Difference of rewards.

$$= \frac{1}{1+\exp(\cdot)}$$

$$P(\xi_A > \xi_B) = \sigma\left(R(\xi_A) - R(\xi_B)\right)$$

$$R(\xi_A) \rightarrow \sum_{t=0}^{T-1} R_\theta(s_t, a_t)$$

GOAL: LEARN $R_\theta(s,a)$

$$\mathcal{L}(\theta) = -\log P(\xi_A \succ \xi_B)$$

$$= -\log \sigma\left(R_\theta(\xi_A) - R_\theta(\xi_B)\right)$$

$$R_\theta(\xi_A) \rightarrow \sum_{t=0}^{T-1} R_\theta(s_t, a_t)$$

RLHF FROM PREFERENCES

1. COLLECT PREF DATASET
2. TRAIN REWARD FUNCTION R_θ
3. CALL RL ALGORITHM R_θ