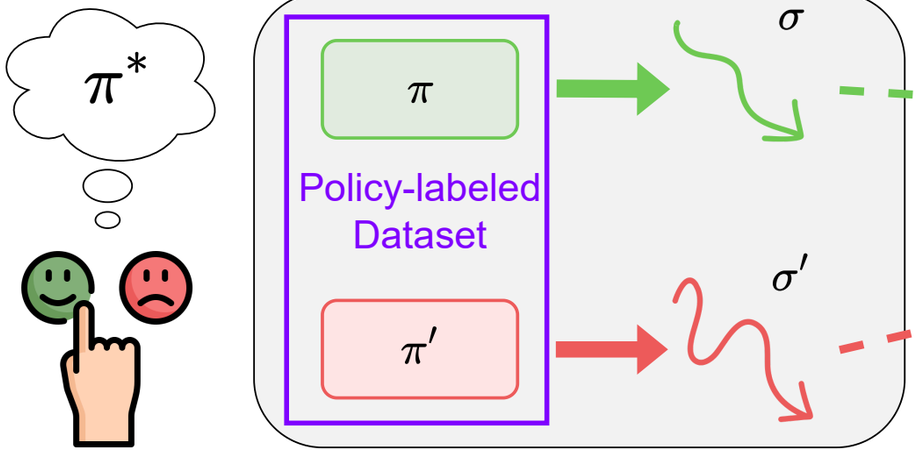
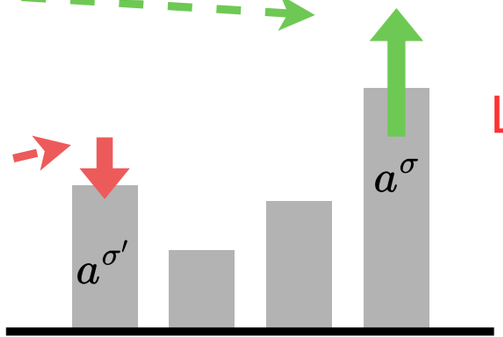
Policy-labeled Preference Learning



$$\sum_{\sigma}\left(\log\pi^*(a_t|s_t)+\mathcal{H}^{\pi}(\cdot|s_t)-\mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{(s_t,a_t)}}\left[\sum_{l>0}\gamma^l D_{KL}(\pi(\cdot|s_l)||\pi^*(\cdot|s_l))\right]\right)$$



Likelihood matched