Explaining Demonstrated Actions

### Learning the Internal Dynamics Model

Our learning algorithm

\[
\text{Demonstrator's internal dynamics model}
\]

\[
\text{Demonstrations}
\]

In an MDP with a discrete action space \( \mathcal{A} \), the human demonstrator is assumed to follow a policy \( \pi \) that maximizes an entropy-regularized reward \( R(s,a,s') \) under dynamics \( T(s',s,a) \). Equivalently,

\[
\pi(a|s) \triangleq \frac{\exp(Q(s,a))}{\sum_{a' \in \mathcal{A}} \exp(Q(s,a'))},
\]

where \( Q \) is the soft Q function, which satisfies the soft Bellman equation,

\[
Q(s,a) = \mathbb{E}_{\pi,T}[R(s,a,s') + \gamma V(s')],
\]

with \( V \) the soft value function,

\[
V(s) \triangleq \log \left( \sum_{a \in \mathcal{A}} \exp (Q(s,a)) \right).
\]

Soft Bellman error:

\[
\delta(s,a) \triangleq Q_i(s,a) - \int_{s' \in \mathcal{S}} T(s',s,a)(R(s,a,s') + \gamma V(s')) \, ds'.
\]

Constrained optimization problem:

\[
\begin{align*}
\min_{\theta} & \sum_{i=1}^n \sum_{s \in S} \sum_{a \in \mathcal{A}} \log \pi_k(a|s) \\
\text{subject to} & \delta_{\theta_k}(s,a) = 0 \quad \forall \, s \in \mathcal{S}, a \in \mathcal{A}.
\end{align*}
\]

Loss function for unconstrained optimization problem:

\[
c(\theta, \phi) \triangleq \sum_{i=1}^n \sum_{s \in S} \sum_{a \in \mathcal{A}} \log \pi_k(a|s) + \frac{1}{2} \sum_{a \in \mathcal{A}} \sum_{s',s \in S} \delta_{\theta_k}(s,a) \lVert \phi \rVert ds. \]

### User Study

We asked 12 participants to play the Lunar Lander game without and with internal-to-real dynamics transfer assistance.

### Likelihood

Videos, code, and data available at [https://sites.google.com/view/inferring-internal-dynamics](https://sites.google.com/view/inferring-internal-dynamics)