**Motivation**

Common forms of exploration in reinforcement learning (e.g., epsilon-greedy and entropy regularization) are undirected. We need smarter, more effective exploration strategies to deal with sparse rewards in high-dimensional action spaces.

**Problem**

- Optimize a policy \( \pi(a) \) over action sequences to maximize expected reward \( \mathbb{E}_{\pi}[r(a)] \), where \( a \sim \pi(a_1, a_2, \ldots, a_n) \).
- The reward landscape \( r(a) \) is not fully observed.

**UREX**

Standard policy-based approach to maximize expected reward.

\[
\mathcal{O}_{\text{UREX}}(\theta) = \mathbb{E}_{\pi}[r(a)]
\]

- Draw \( K \) i.i.d. action sequences \( \{a^{(k)}\} \) and estimate the current policy, i.e., \( \pi(a^{(k)}) = \pi(a) \) for each \( 1 \leq k \leq K \).
- Estimate the gradient using importance sampling.

**Characteristics of UREX**

- Rather than undirected exploration, UREX encourages exploration in areas where under-estimated rewards are under estimated by the current policy.
- \( \bar{w}^{(k)} \) measures the difference between \( r(a^{(k)})/\tau \) and \( \log \pi(a^{(k)}) \), and normalized importance weights find the most under-appreciated action sequences among \( K \) samples.
- UREX is simple and easy to implement.

**Justification**

- Recall KL divergence between distributions \( p(a) \) and \( q(a) \):

\[
-\text{KL}(p || q) = \mathbb{E}_{p}[\log p(a)] + \mathbb{E}_{q}[\log q(a)]
\]

- We re-express the entropy regularized objective as a KL.

\[
-\text{KL}(\pi \| \bar{\pi}^*) = \int (\pi(a) \log \pi(a) - \log \pi(a^{\tau}))
\]

- \( \bar{\pi}^* \) has a mode seeking behavior, prone to falling into local minima.

- \( \bar{\pi}^{(k)} \) has a mode covering behavior, but requires sampling from \( \pi(a) \).

**MENT**

Augment the objective with entropy regularization.

\[
\mathcal{O}_{\text{MENT}}(\theta, \tau) = \mathbb{E}_{\pi}[r(a)/\tau] + \mathbb{E}_{\pi}[\log \pi(a)]
\]

- Estimate the gradient using \( K \) on-policy samples.

**Results**

- UREX reliably solves reversion and multi-digit addition.
- UREX \( \geq \text{MENT} \geq \text{REINFORCE} \).

The RL agents only observe total reward at the end of episode.

**Future Directions**

- Make use of rewards per time step.
- Exploit off-policy trajectories.
- Exploit expert trajectories.
- Combine with trust region methods.

**Softmax Temporal Consistency**


Builds on top of this work to exploit per-step rewards. Proposes a softmax temporal consistency between a state-action pair \((s, a)\) and a subsequent state \(s'\):

\[
Q'(s, a) = r(s, a) + \gamma \log \sum_a \exp(Q(s', a')/\tau)
\]

**References**

- See Williams, et al. (1991) and Williams (1992) for more on REINFORCE/MENT.
- See Norouzi, et al. (2016) for the use of mode covering KL divergence for structured output prediction.
- Our paper discusses the similarities and differences of UREX with Reward-Weighted Regression (Peters & Schaal, 2007).