PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Samuel Ainsworth, Matt Barnes, Siddhartha Srinivasa

Mo' States Mo' Problems: Emergency Stop Mechanisms from Observation

SUMMARY

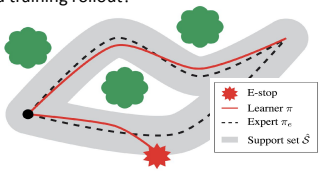
In many RL environments, optimal policies only visit a small subset of the state space. "Emergency stops" exploit this phenomenon, revealing a tradeoff between learner sample complexity and asymptotic performance.

1. Emergency stops

Key question: When should an RL agent be stopped along a training rollout?

Emergency stops (e-stop): Premature ends to rollouts.
Manual e-stop: Human with a big red button.
Automatic e-stop: Learned termination condition.

Support set $\hat{\mathcal{S}}$: Set of states which do not induce e-stops.



2. The sample complexity vs. suboptimality tradeoff

$$\text{Regret}(T) \leq \underbrace{\left\lceil \frac{T}{H} \right\rceil [J(\pi_e) - J(\hat{\pi}^*)]}_{\text{Asymptotic sub-optimality}} + \underbrace{\mathbb{E}^{\hat{\pi}^*}_{\widehat{M}}[R_T] - \mathbb{E}^{\mathfrak{A}}_{\widehat{M}}[R_T]}_{\text{Learning regret}}$$

3. Guarantees

Learning Regret scales with the size of the state space, e.g. well known lower bound $\Omega(\sqrt{HSAT})$

This suggests a fundamental tradeoff between

- The frequency of e-stop, via the size of $\hat{\mathcal{S}}$
- The performance of the agent

Asymptotic suboptimality is bounded in terms of the expert's average state distribution.

Corollary 4.2.1. *Recall that $\rho_{\pi_e}(s)$ denotes the average state distribution following actions from π_e, $\rho_{\pi_e}(s) = \frac{1}{H}\sum_{t=0}^{H-1} \rho^t_{\pi_e}(s)$. Then*

$$J(\pi_e) - J(\hat{\pi}^*) \leq \rho_{\pi_e}(\mathcal{S} \setminus \hat{\mathcal{S}})H^2 \qquad (7)$$

State-only expert demonstrations can be used to construct the support superset $\hat{\mathcal{S}}$

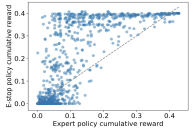
Theorem 5.1. *The e-stop MDP $\widehat{M}$ with states $\hat{\mathcal{S}}$ in Algorithm 1 has asymptotic sub-optimality*

$$J(\pi_e) - J(\hat{\pi}^*) \leq (\xi + \epsilon)H \qquad (8)$$

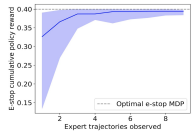
with probability at least $1 - |\mathcal{S}|e^{-2\epsilon^2 n/|\mathcal{S}|^2}$, for any $\epsilon > 0$. Here ξ denotes our approximate state removal "allowance", where we satisfy $\sum_{s \in \mathcal{S}\setminus\hat{\mathcal{S}}} \hat{h}(s) \leq \xi$ in our construction of $\widehat{M}$ as in Theorem 4.2.

Implications: Remove states that are low probability under the expert's average state distribution.

4. Experiments



E-stop performance tends to exceed expert performance.



The support set can be constructed out of a surprisingly small number of expert demonstrations.

E-stops empirically trade off a small suboptimality gap for a significant increase in sample complexity.
- Support sets built from pre-trained expert policy roll-outs.