Working Group “Probabilistic and Interactive Machine Learning”

AAMAS’23: Learning Constraints From Human Stop-Feedback

Sebastian Tschiatschek (with S. Poletti and A. Testolin)

Abstract

- Safety: central for the usage of intelligent agents in many domains
- In this paper: learning about dangerous behavior via stop-feedback in RL
  - Probabilistic feedback model inspired by how humans might provide feedback
  - Bayesian inference for inferring constraints
- Experiments:

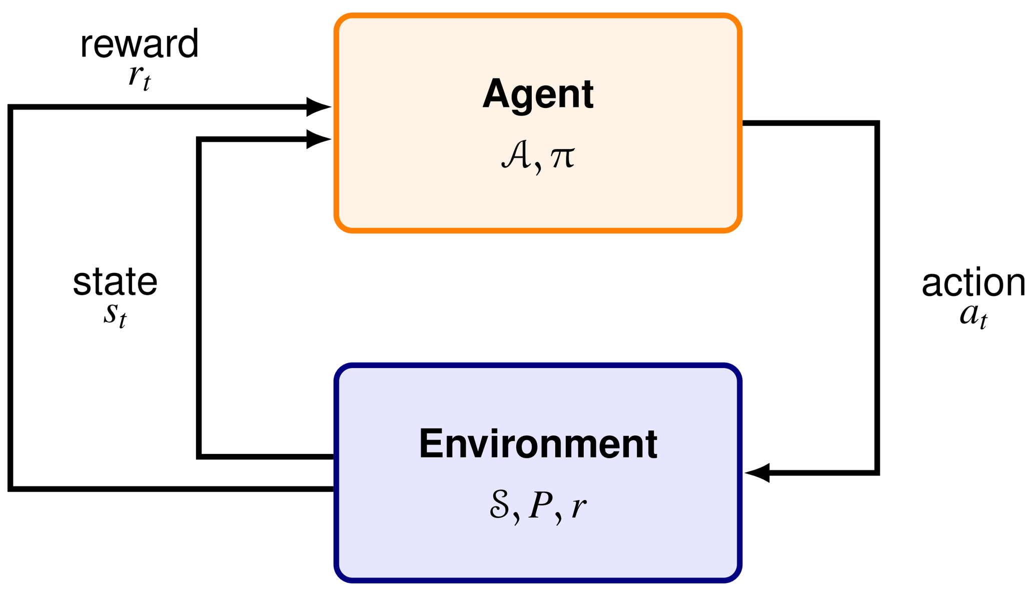
  Learning with our proposed feedback model is efficient

  Human stop-feedback aligns reasonably well with our model

Setting

Markov Decision Processes

$$\mathcal{M} = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$$



Classical Goal

- Cumulative rewards:

$$J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t) \mid \pi\Big],$$

- Find optimal policy:

$$\pi^* = \arg\max_{\pi} J(\pi).$$

Here: Learning With Constraints

- Rewards $r\colon \mathcal{S} \to \mathbb{R}_+$ known
- Constraints $c\colon \mathcal{S} \to \mathbb{R}_+$ unknown
- Dangerous states $\Leftrightarrow c(s) \gg$
- Agent should avoid dangerous states
- Find optimal policy for

$$J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t) - c(s_t)\big) \mid \pi\Big],$$

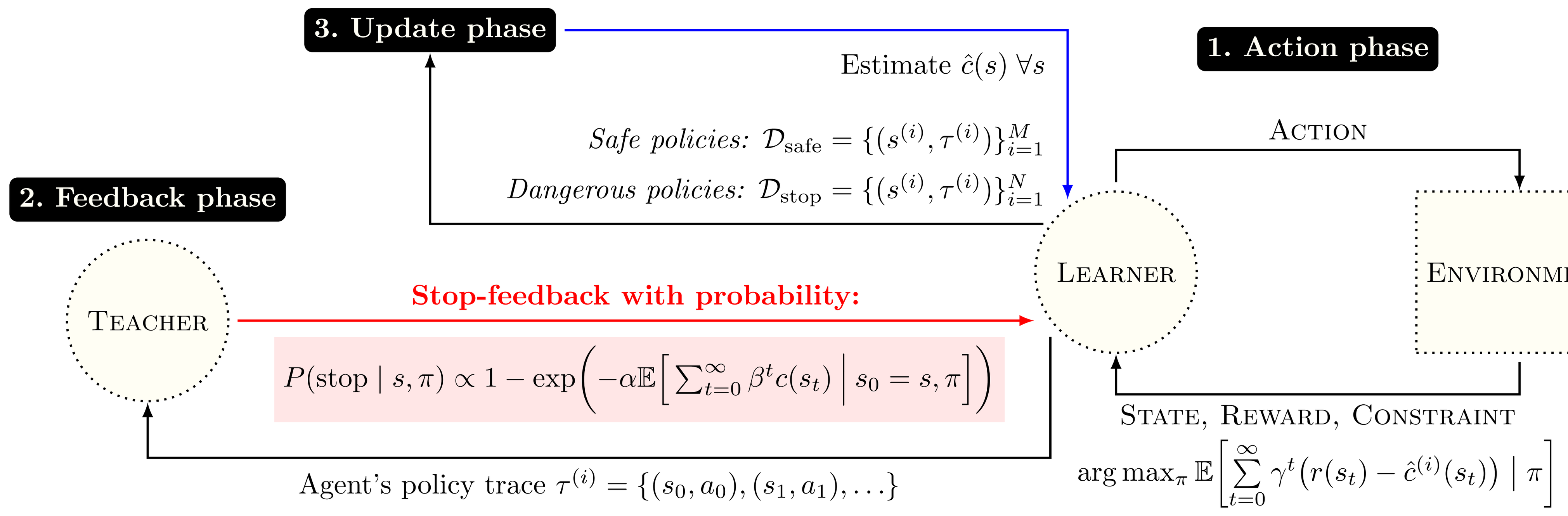
Stop-Feedback

- Learn about constraints from stop-feedback (provided by a human supervisor)
- Model how a human supervisor might provide feedback

$$P(\text{stop} \mid s, \pi) \propto 1 - \exp\left(-\alpha \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t c(s_t) \,\middle|\, s_0 = s, \pi\right]\right) \quad (1)$$

  - β: Horizon for reasoning into the future
  - α: ”Worriedness” of supervisor
- Goal: sample-efficient learning from stop-feedback

Interaction of Agent and Supervisor



Our Approach

Central Ideas

- Estimate constraints from stop-feedback
- Be Bayesian:
  - Encourage exploration
  - Incorporate prior knowledge
- Featurize environment for scalability:

$$c(s) = \langle \phi(s), c^* \rangle$$

Major Steps

1 *Action phase*

2 *Feedback phase*

3 *Update phase*

- Compute $\hat{c}^{(i+1)}$ as posterior via MCMC
- Optimize policy:

$$\pi^{(i)} = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \big(r(s_t) - \hat{c}^{(i)}(s_t)\big) \middle| \pi\right] \quad (2)$$

```
Input: maximum number of interactions K
Output: Final learner's policy π^{K+1}
/* Initialization */
D_stop ← ∅, D_safe ← ∅
ĉ^{(1)}(s) ← 0   ∀s ∈ S
π^{(1)} ← (approx.) optimal policy for r, ĉ^{(1)}, cf. Eq. (2)
/* Learner-teacher interaction */
for all i = 1, ..., K do
    /* Action & feedback phase */
    s ← s_0
    for all t = 1, ..., T do
        a_t ~ π^{(i)}(s), s_{t+1} ~ P(· | s_t, a_t), r_t ~ r_{a_t}(s_t)
        f ← Teacher's feedback according to Eq. (1)
        if f = stop then
            D_stop ← D_stop ∪ {(s_t, π^{(i)})}
            break
        else
            D_safe ← D_safe ∪ {(s_t, π^{(i)})}
    /* Update phase */
    Learner updates its estimate of the constraints to
    ĉ^{(i+1)} based on the datasets D_stop, D_safe
    π^{(i+1)} ← (appr.) optimal policy for r, ĉ^{(i+1)}, cf. Eq. (2)
return Final learner's policy π^{K+1}
```

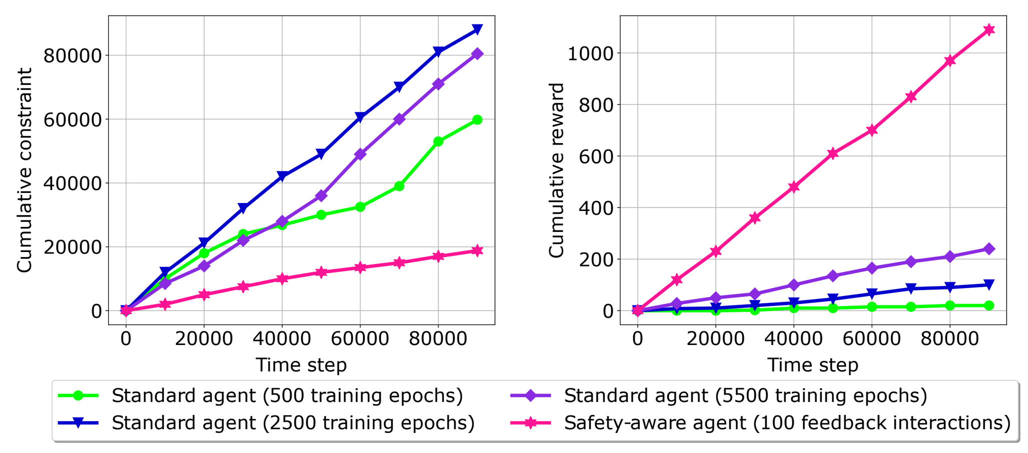
Experimental Setup

OpenAI Safety Gym

- 1 goal state
- 5 evenly distributed fixed hazards
- Agent can move in the 2d plane (turning & moving forward/backward)
- Environment is reset when reaching hazard
- PPO; 2$^{\text{nd}}$ layer of critic for feature extraction

Experimental Results

Learning from Synthetic Feedback

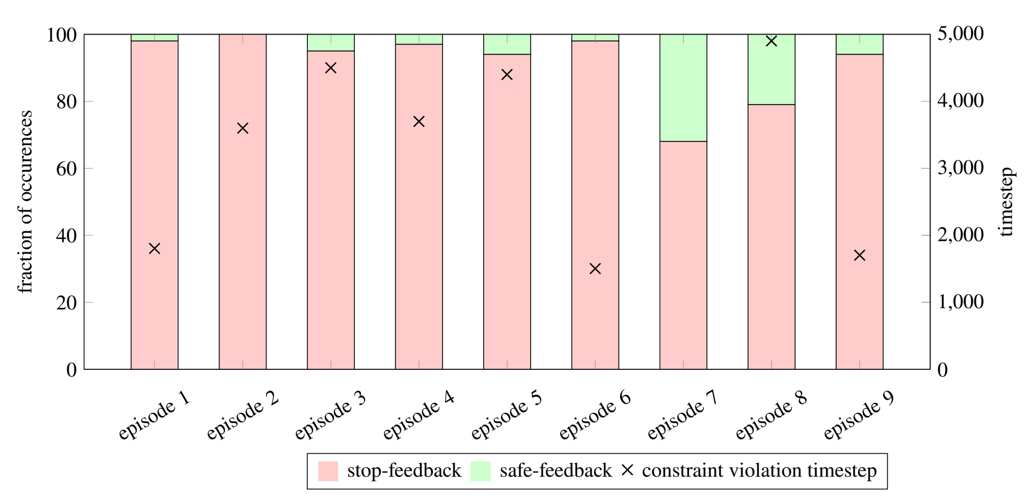


- Agents with constraint inference:
  - Achieve higher cumulative rewards
  - Violate fewer constraints

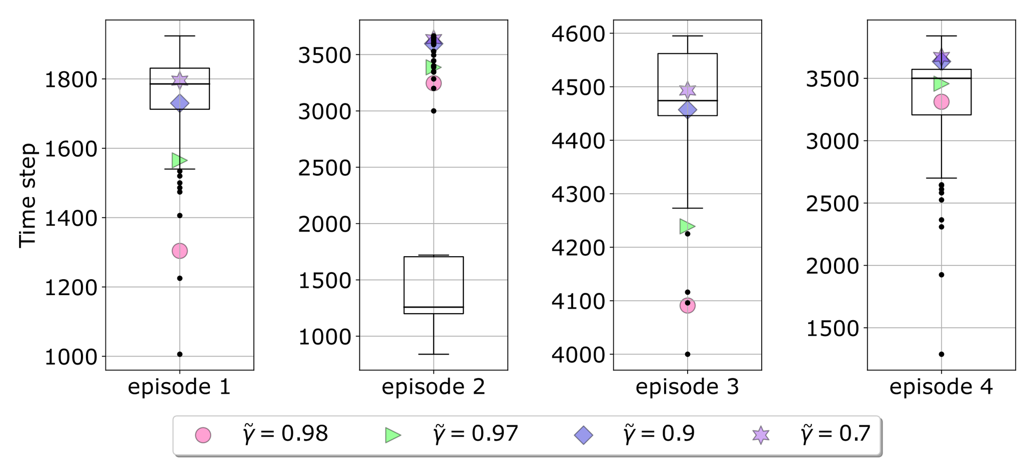
Human Feedback

- Survey with 100 volunteers (online)
- Evaluation of 9 videos:
  - Episodes with 5000 time steps
  - *Standard agent* colliding with $\leqslant 1$ constraint
- Question:
  - Stop-feedback or not
  - Time step for feedback

Stop-feedback and collision times



Model-generated vs human stop-feedback



- Good alignment of feedback

Further Details

Paper link

Group link