High Confidence Generalization for Reinforcement Learning

IoBT REIGN

Relevant to IOBT CRA Research Area / Task: Research Area 2, Task 2.3

Objectives

- Develop high confidence generalization algorithms (HCGAs) for reinforcement learning (RL) that are adaptive, can solve a wide variety of tasks, and are robust to changing conditions and situations.
- Draw on principles from safe (Seldonian) machine learning to create a class of reinforcement learning (RL) algorithms that generalize safely across a distribution of tasks.

Approach

- The user 1) defines a **minimum safe measure of performance** (in terms of expected return) for the RL task distribution and 2) specifies **the probability with which the algorithm must achieve this measure of performance** on Markov decision processes (MDPs) drawn from this distribution. The algorithm trains on MDPs sampled from a distribution of tasks, and returns either: 1) a solution that is guaranteed to satisfy the safety constraint with the specified probability, or 2) if a safe solution cannot be found with the specified probability, the algorithm returns “No Solution Found”. By modifying the types of bounds we use, we can guarantee performance even when the target task does not come from the training distribution (assuming other assumptions are met, and that sufficient data exists).

Input:

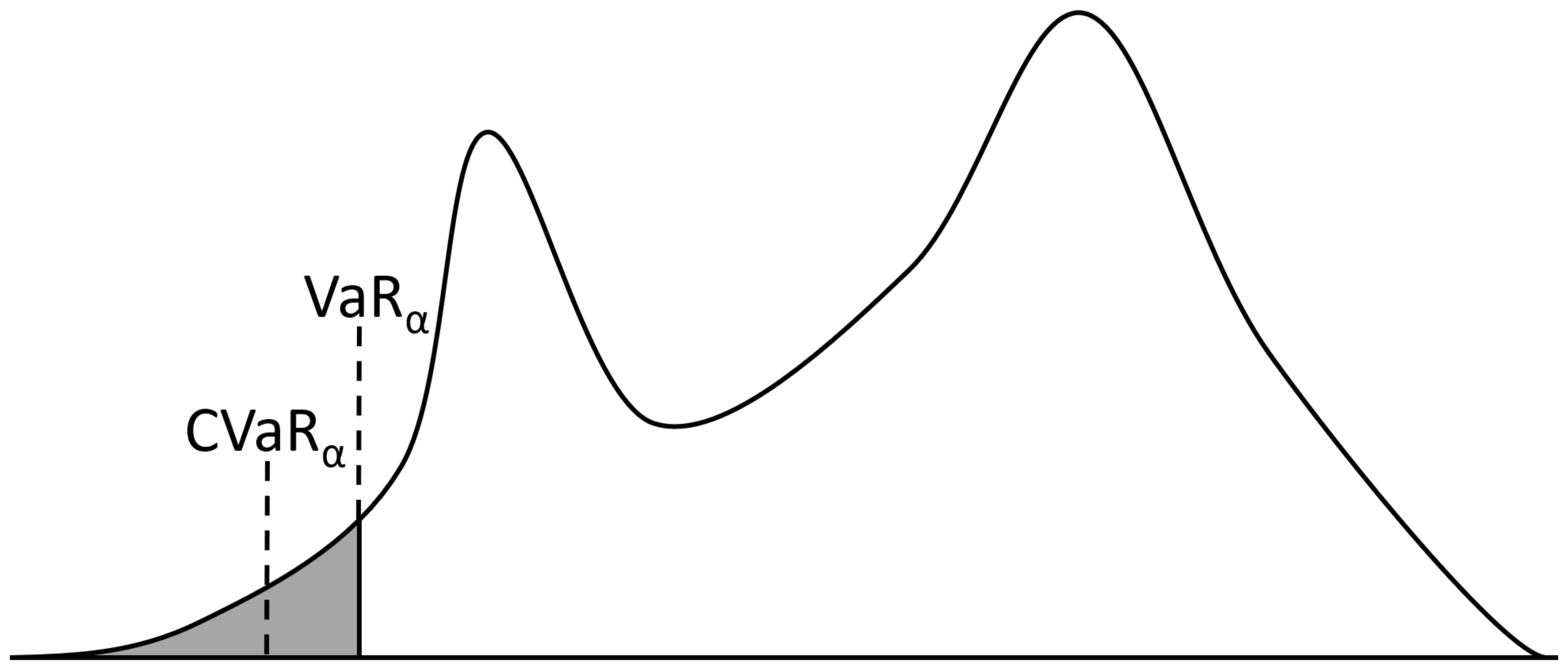
- Feasible set Θ, a set of MDPs M, and probability $1-\delta$.
- Objective function J such that $J(\theta, M) \in [0,1]$ is an estimate of the utility of the solution θ, computed using data M.
- A satisfactory expected return c that our algorithm should achieve with at least probability $1-\delta$.

Output: A solution, $\theta \in \Theta$, or NO SOLUTION FOUND.
Partition M into two data sets, M_1 and M_2;
$\theta_c = \text{argmax}_{\theta\in\Theta} J(\theta, M_1)$;
if $J(\theta_c, M_2) - \sqrt{\frac{\ln(1/\delta)}{2|M_2|}} \geq c$ then return θ_c;
return NO SOLUTION FOUND;

Pseudocode of a simple example algorithm in this class, using bounds based on Hoeffding's inequality.

Results

- We have implemented several algorithms in this class. There are two variants for expected value, and four variants based on risk measures.
- Risk sensitive measures:



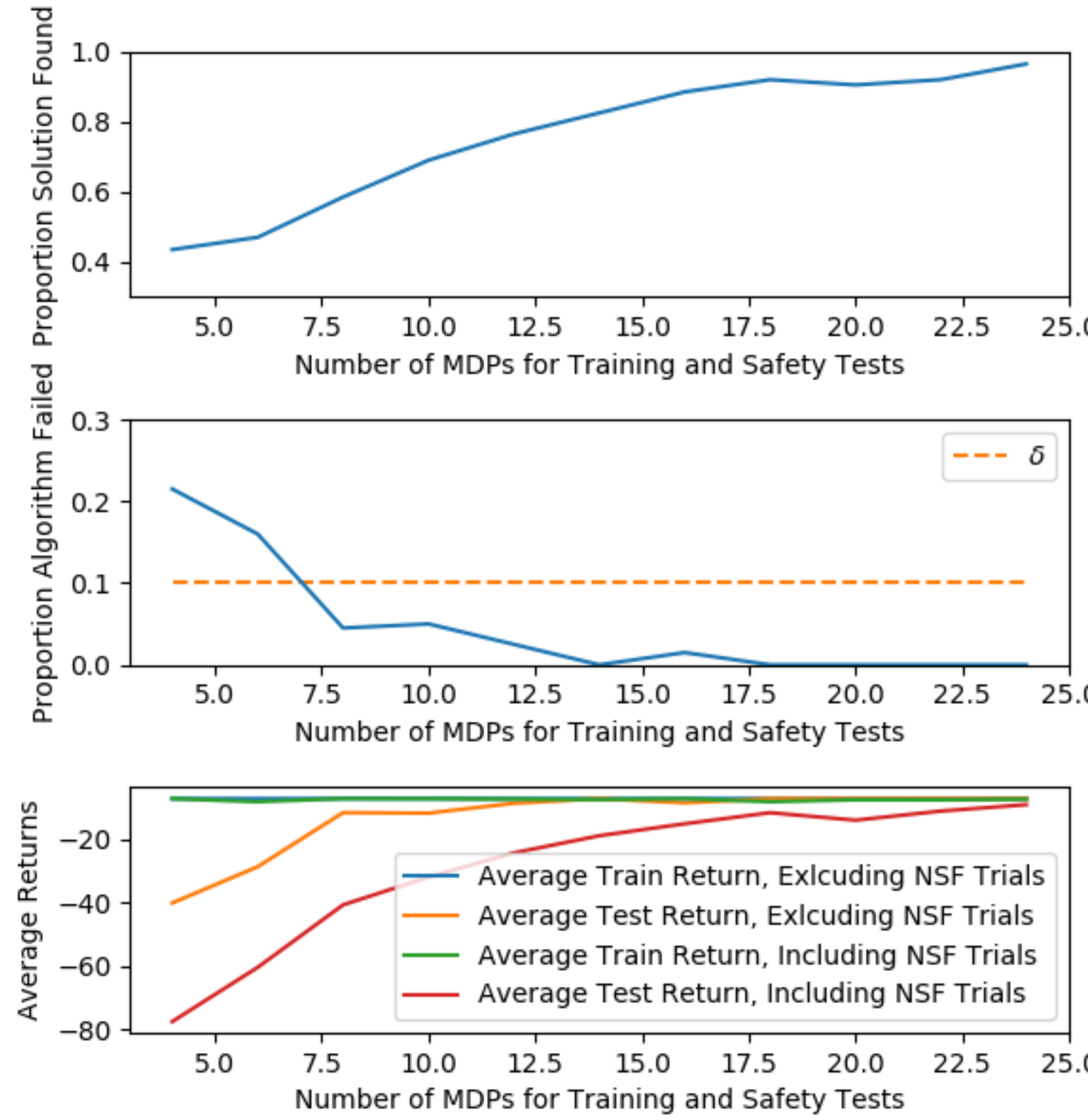
- Instead of expected performance on the distribution of tasks, we can use percentile-based risk measures (VaR and CVaR) to specify performance on MDPs drawn from the distribution. Intuition: safety constraints on the “worst case” MDPs in the distribution.

$$\Pr(\text{VaR}_\alpha(J_{M_1}(a(M_{\text{acc}}))|M_1 \sim \mu, M_{\text{acc}} \sim \mu) \geq j) \geq 1-\delta,$$
$$\Pr(\text{CVaR}_\alpha(J_{M_1}(a(M_{\text{acc}}))|M_1 \sim \mu, M_{\text{acc}} \sim \mu) \geq j) \geq 1-\delta,$$

- Another class of interest is episodic bounds: instead of guaranteeing average performance, we can apply percentile-based risk measures (VaR and CVaR) to the distribution of episodes. Intuition: safety constraints on the “worst case” episodes.

$$\Pr(\text{VaR}_\alpha(G_{M_1}(a(M_{\text{acc}}))|M_1 \sim \mu, M_{\text{acc}} \sim \mu) \geq j) \geq 1-\delta,$$
$$\Pr(\text{CVaR}_\alpha(G_{M_1}(a(M_{\text{acc}}))|M_1 \sim \mu, M_{\text{acc}} \sim \mu) \geq j) \geq 1-\delta,$$

- The results below come from an HCGA variant that uses Student’s t-test for the bound, and an actor-critic algorithm for optimization.



Conclusions

These algorithms are adaptive, safe, and they work: we’ve laid the theoretical foundations, and initial experiments show that the safety guarantees hold in practice.

Path Forward

- Continue to develop related classes of safety guarantees.
- Gather more empirical results on different kinds of tasks.
- Work on the optimal stopping problem represented by how to split the training and safety data sets.
- Extend the algorithms, in theory and in practice, to function in the extrapolation setting.

ILLINOIS · USC University of Southern California · UMassAmherst · UCLA · SRI International · Georgetown University

POINT OF CONTACT:
James Kostas
jekostas@umass.edu