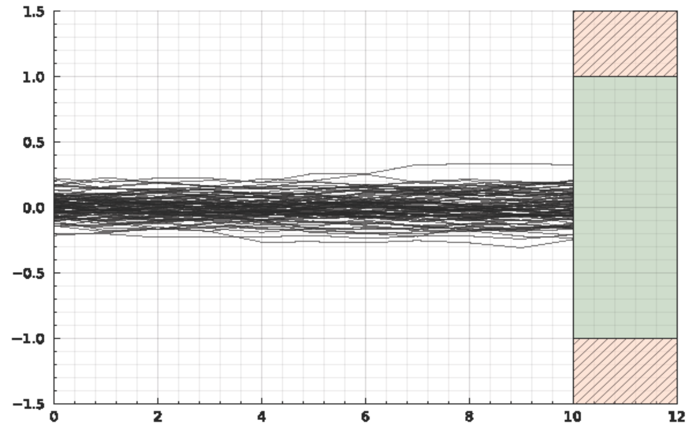
Discovery and Analysis of Rare High-Impact Failure Modes Using Adversarial RL-Informed Sampling

Rory Lipkis and Adrian Agogino

Motivation

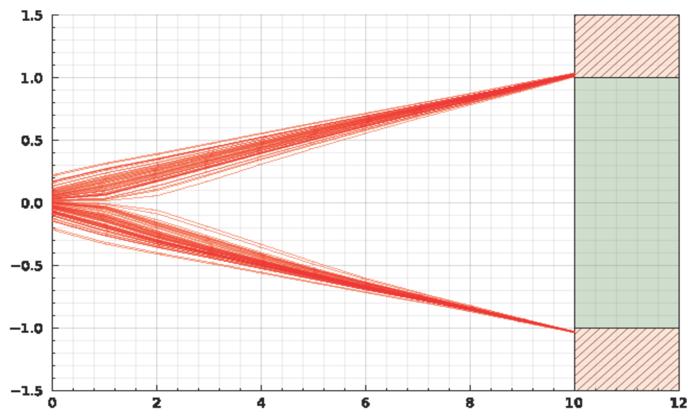
- Complex autonomous systems require verification and validation
- Rare failures are difficult to find; even more difficult to analyze
- Monte Carlo sampling can be inefficient, intractable, and misleading
- Common acceleration strategies introduce "expert" bias and jeopardize independent testing

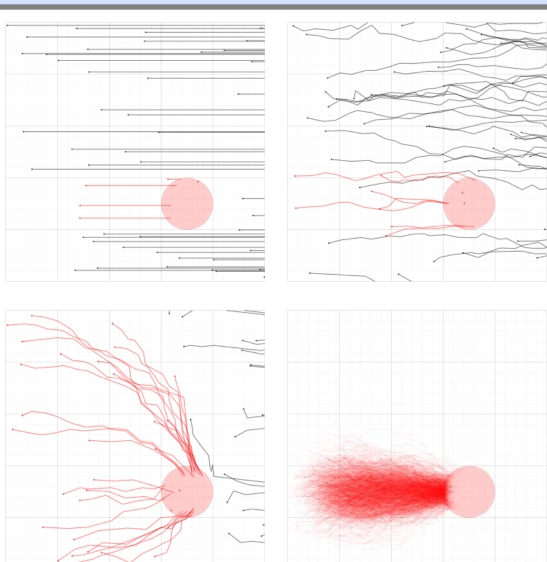


Formalization

- System state s
- Stochastic time-varying environment $X \sim p(x)$
- Failure criterion $s \in F$
- System evolution $s_{t+1} = s_t(x)$
- True failure probability:

$$\mu = E_{\mathbf{x}}[\mathbf{1}_F(s(\mathbf{x}))]$$





Method

- Learn failure policy $\pi^*(s)$ that encodes statistical modes of failure-conditional distribution
- Form surrogate environment $q(x)$ based on $\pi^*(s)$
- Importance resampling:

$$\mu \approx \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_F(s(\mathbf{x}^{(i)}))\frac{p(\mathbf{x}^{(i)})}{q(\mathbf{x}^{(i)})}$$