Sandia National Laboratories

Data-Driven Control Strategies for PDE Environments using Reinforcement Learning

Nick Winovich†, Deepanshu Verma‡, Lars Ruthotto‡, and Bart van Bloemen Waanders†

Research Goals

- develop **global** strategies for controlling complex physical systems
- provide actionable control information in real-time
- identify a general framework adaptable to diverse system dynamics and flexible enough to account for practical constraints on actions

Challenges

- reliance on computationally demanding forward models
- limited number of data requests available due to complexity
- problem parameters and input conditions are unknown until the event commences and an immediate response is required
- infinite dimensional space of system configurations and policies

Model Environment for Reinforcement Learning

PDE System

$$\frac{\partial u}{\partial t} + \mathbf{v}_\phi \cdot \nabla u - D \cdot \Delta u \;=\; f_{\xi,\omega} - a \quad \text{in} \quad \Omega \times [0,T] \quad \text{with} \quad D = 0.5$$

$$u = 0 \;\text{ on }\; \{x=0\} \cup \{y=0\} \cup \{y=1\} \quad \text{and} \quad \frac{\partial u}{\partial n} = 0 \;\text{ on }\; \{x=1\}$$

Source Term and Velocity Field

$$f_{\xi,\omega}(x,y) \;=\; 5.0/\sigma \cdot \exp\left(-(|x-\xi| + |y-\omega|)/\sigma\right) \quad \text{with} \quad \sigma = 0.01$$

where $\xi \sim$ **Uniform(0.1,0.25)** and $\omega \sim$ **Uniform(0.1,0.9)**

$$\mathbf{v}_\phi(x,y) \;=\; \left(\sqrt{\eta^2 - \delta^2 \cdot \sin^2(2\pi \cdot [x-\phi])}\;,\; -\eta \cdot \sin(2\pi \cdot [x-\phi]) \right)$$

where $\eta = 12.5$, $\delta = 0.75$, and $\phi \sim$ **Uniform(0.0,-0.3)**

Control Decision

Select an action $A_t = [r_t, v_t]$ for adjusting the initial magnitude $M_0 = 0.0$ and initial position $P_0 = 0.5$ of the sink expression $a(x,y,t)$ given by:

$$a(x,y,t) \;=\; M_t \cdot \exp\left(-(|x-0.6|/\sigma_x + |y-P_t|/\sigma_y)\right)$$

$$M_{t+1} \;=\; M_t + r_t \cdot \Delta t \quad \text{and} \quad P_{t+1} \;=\; P_t + v_t \cdot \Delta t$$

Objective function for model environment

$$\min_{\{r_t, v_t\}} \; \mathbb{E}_{\xi,\omega,\phi} \left[\underbrace{\int_{\Omega_T} |u(x,y,T)|^+ \, dm}_{\text{Outcome of Event}} + \lambda \cdot \underbrace{\int_0^T |r(t)|^2 + |v(t)|^2 \, dt}_{\text{Cost to Act}} \right]$$

Reinforcement Learning with Actor-Critic Models

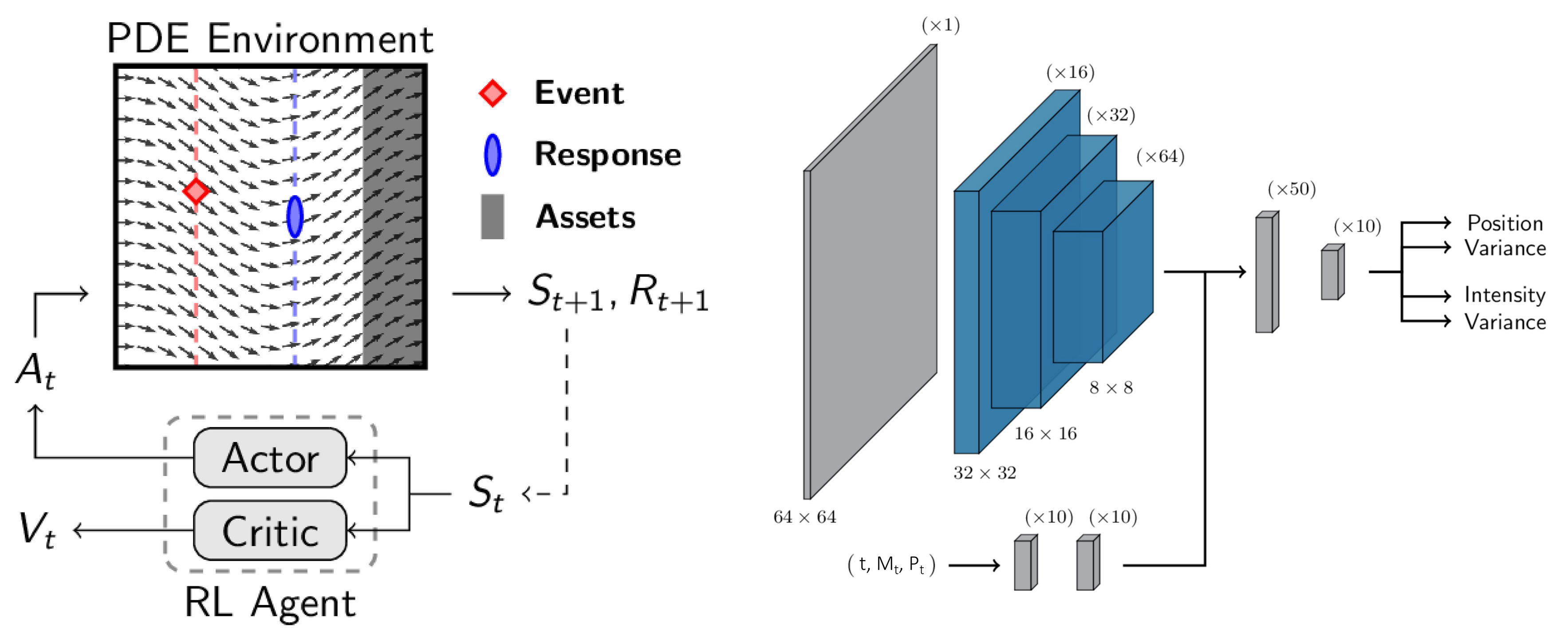
- an *actor* network is tasked with proposing control actions at each time-step based on the current system state
- a *critic* network is trained to predict the long-term value/outcome of the system based on the current state of the environment and actor
- the actor must refine its decisions to outperform the critic's prediction

Proximal Policy Optimization (PPO)

Schulman, John, et al. "Proximal policy optimization algorithms." *arXiv preprint arXiv:1707.06347* (2017).

- avoid over-tuning during training using trust-regions to select step-size
- move cautiously if feedback is positive, move decisively if negative

Training workflow and neural network architecture

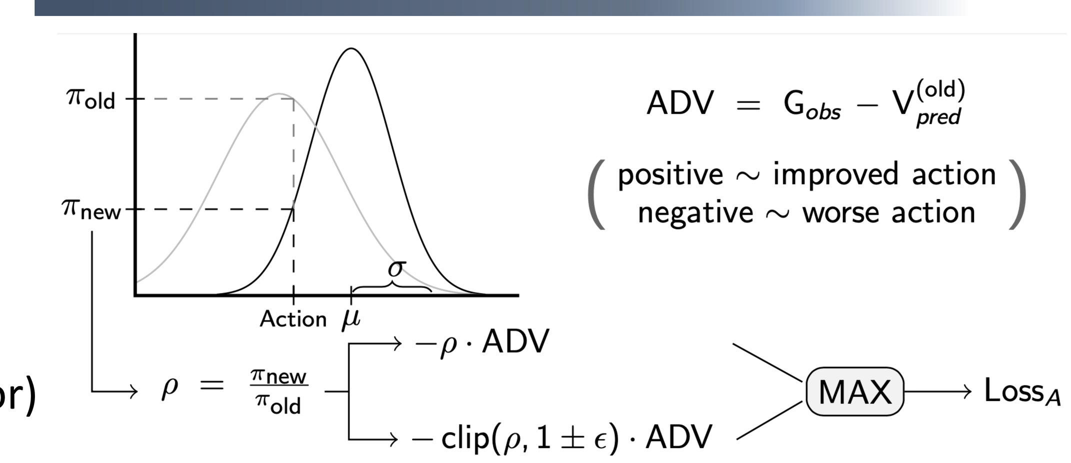


The RL agent is trained through repeated interactions with various environment realizations by:

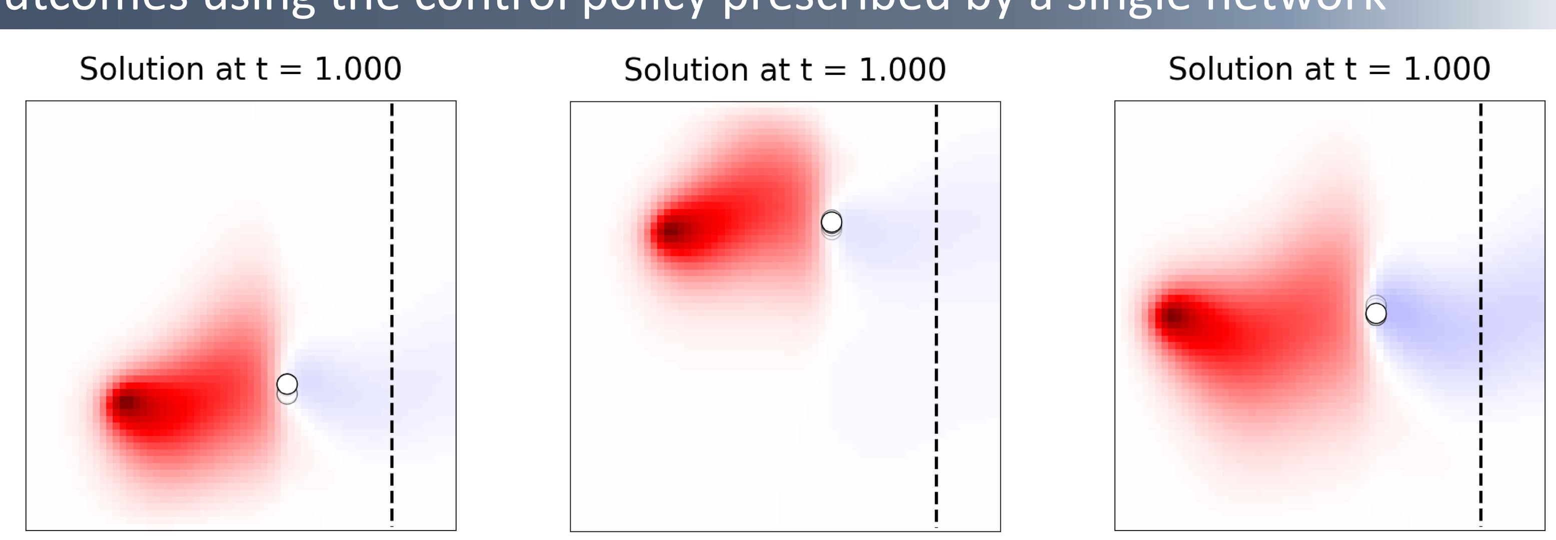
1) *estimating the value* $\mathbf{V}_t$ of the current system state $\mathbf{S}_t$ (critic)

2) *proposing an optimal course of action* $\mathbf{A}_t$ at each time step (actor)

Actor loss for PPO



Outcomes using the control policy prescribed by a single network

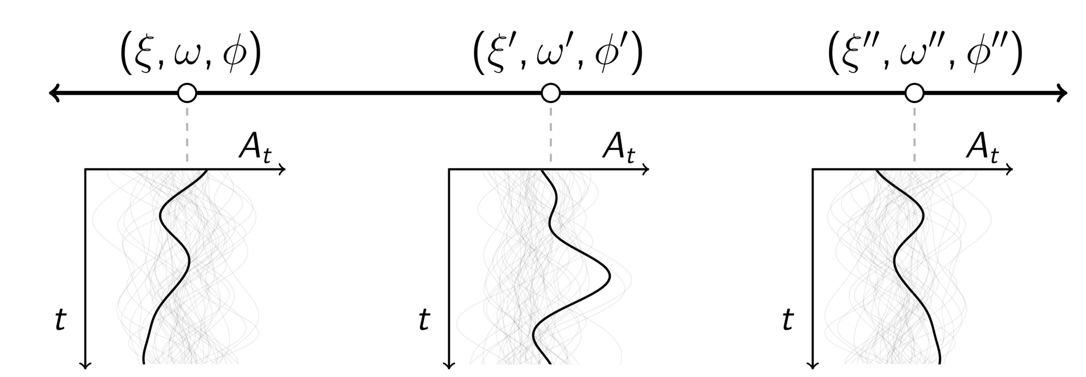


A single trained agent can quickly produce effective containment strategies for a variety of distinct problem realizations without referencing the forward model.

Search Space Complexity

- 3 continuously varying environment parameters which each have a notable effect on the overall system dynamics
  - ➢ unlikely to see the same realization more than once
- 25 actions must be selected sequentially for each realization
  - ➢ curse of dimensionality as number of time-steps increases
- search is performed using the scalar-valued objective alone
  - ➢ no gradient information or system knowledge

Local versus Global Solutions



- *local methods* apply to a single realization of the system and require repeated simulation calls once parameters are known
  - ➢ provide a solution for one specific set of parameter values
- *global/semi-global methods* are calibrated offline using simulation data reflecting a broad range of system realizations
  - ➢ yield approximate solutions for a distribution of parameters

Key Takeaways

- RL successfully navigates the infinite dimensional search space using a finite sequence of forward model queries
- minimal run-time costs and no additional model queries
- framework is applicable to a diverse family of problems
- flexible implementation, model treated as a black-box
- data inefficient due to lack of system specific knowledge

Future Work

- incorporate physical knowledge of system into training
- take advantage of the mathematical structure prescribed by the FEM-discretized weak formulation
- enforce constraints on the actor-critic networks dictated by the Hamilton-Jacobi-Bellman equations

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology &Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
SAND No. 2022-4391 C

† Sandia National Laboratories ‡ Emory University