Actor-Critic Methods using Physics-Informed Neural Networks: Control of a 1D PDE Model for Fluid-Cooled Battery Packs

UNIVERSITY OF WATERLOO

ICML International Conference On Machine Learning

New Frontiers in
Learning, Control, and Dynamical Systems

Amartya Mukherjee, Jun Liu

Department of Applied Mathematics, University of Waterloo

Abstract

This paper proposes an actor-critic algorithm for controlling the temperature of a battery pack using a cooling fluid. This is modeled by a coupled 1D partial differential equation (PDE) with a controlled advection term that determines the speed of the cooling fluid. The Hamilton-Jacobi-Bellman (HJB) equation is a PDE that evaluates the optimality of the value function and determines an optimal controller.

- We propose an algorithm that treats the value network as a Physics-Informed Neural Network (PINN) to solve the continuous-time HJB equation
- We derive a control function from the HJB equation.

Our experiments show that a hybrid-policy method that updates the value network using the HJB equation and updates the policy network identically to PPO achieves the best results in the control of this PDE system.

The 1D pack cooling problem

Modelled by the following coupled PDE

$$u_t(x,t) = -D(x,t)u_{xx}(x,t) + h(x,t,u(x,t)) + \frac{1}{R(x,t)}(u-w)$$
$$w_t(x,t) = -\sigma(t)w_x(x,t) + \frac{1}{R(x,t)}(w-u)$$

with the following boundary conditions

$$u_x(0,t) = u_x(1,t) = 0$$
$$w(0,t) = U(t)$$

where

u: Temperature distribution across battery pack
w: Temperature distribution across cooling fluid
D: Thermal diffusion constant
R_T: Thermal resistance
h: Internal heat generation in the battery pack
U: temperature of the cooling fluid at the boundary
σ: Transport speed of the cooling fluid (controller)

Objective of the controller:

Maximize $\int_0^\infty \int_0^1 \gamma^t (-u(x,t)^2)\,dx\,dt$

Environment parameters:

$$h(x,t,u(x,t)) = e^{0.1u(x,t)}$$
$$u(x,0) = \sum_{n=0}^{9} C_n \cos(\pi n x)$$
$$\Delta x = 0.01, \Delta t = 0.01$$
$$U(t) = -5.0, D(x,t) = 0.01, R(x,t) = 2.0$$

The HJB equation

Consider a controlled dynamical system modeled by the following equation:

$$\dot{x} = f(x,u), \qquad x(t_0) = x_0$$

In control theory, the optimal value function $V^*(x)$ is useful towards finding a solution to control problems:

$$V^*(t) = \sup_{\sigma} \frac{1}{\Delta t}\int_{t_0}^{\infty} \gamma^{\frac{t}{\Delta t}} L(x(\tau; t_0, x_0, \sigma(\cdot)), \sigma(\tau))\,d\tau$$

where $L(x,\sigma)$ is the reward function, Δt is the time step size for numerical simulation, and γ is the discount factor.

Theorem 2.1. A function $V(x)$ is the optimal value function if and only if:

1. $V \in C^1(\mathbb{R}^n)$ and V satisfies the Hamilton-Jacobi-Bellman (HJB) Equation

$$(\gamma-1)V(x) + sup_{\sigma\in U}\{L(x,\sigma) + \gamma\Delta t\nabla_x V^T(x) f(x,\sigma)\} = 0$$

2. For all $x \in \mathbb{R}^n$, there exists a controller $\sigma^*(\cdot)$ such that:

$$L(x,\sigma^*(x)) + \gamma\Delta t\nabla_x V^T(x) f(x,\sigma^*(x)) = \sup_{\hat{\sigma}}\{L(x,\hat{\sigma}(x)) + \gamma\Delta t\nabla_x V^T(x) f(x,\hat{\sigma}(x))\}$$

Discretize the PDE in space to form an ODE

$$\dot{U} = -DAU + h(U) + \frac{1}{R}(W-U)$$
$$\dot{W} = -\sigma(t)BW + \frac{1}{R}(U-W)$$

where AU approximates u_{xx} and BWapproximates w_x using finite differences.
Use this discretization to derive the HJB equation.

HJB control of the pack cooling problem

Theorem 4.1. Let $u(\cdot,t), w(\cdot,t) \in L_2[0,1]$. With $\sigma(t) \in [0,1]$ and the reward function $L(u(\cdot,t), w(\cdot,t), \sigma(t)) = -||u(\cdot,t+\Delta t)||_2^2$, the HJB equation for the 1D pack cooling problem is:

$$(\gamma-1)V - ||u(\cdot,t+\Delta t)||^2 + \langle V_u(u(\cdot,t),w(\cdot,t)), u_t(\cdot,t)\rangle + \frac{1}{R}\langle V_w(u(\cdot,t),w(\cdot,t)), u(\cdot,t)-w(\cdot,t)\rangle + \max(0, -\langle V_w(u(\cdot,t),w(\cdot,t)), w_x(\cdot,t)\rangle) = 0$$

where $||\cdot||$ is the $L_2[0,1]$ norm and $\langle\cdot,\cdot\rangle$ is the $L_2[0,1]$ inner product.

Corollary 4.2. Let $u(\cdot,t), w(\cdot,t) \in L_2[0,1]$. With $\sigma(t) \in [0,1]$ and the reward function $L(u(\cdot,t), w(\cdot,t), \sigma(t)) = -||u(\cdot,t+\Delta t)||_2^2$, provided the optimal value function $V^*(u,w)$ with $V_w^*(\cdot,t) \in L_2[0,1]$, the optimal controller for the 1D pack cooling problem is:

$$\sigma^*(t) = \begin{cases} 1, \langle V_w^*(u(\cdot,t)), w(\cdot,t)\rangle < 0, \\ 0, otherwise \end{cases}$$

where $\langle\cdot,\cdot\rangle$ is the $L_2[0,1]$ inner product.

New loss functions for the value network

Derived from the HJB equation:

$$MSE_f = \frac{1}{T}\sum_{t=0}^{T-1}\Big((\gamma-1)V - ||u(\cdot,t+\Delta t)||^2 + \langle V_u(u(\cdot,t),w(\cdot,t)), u_t(\cdot,t)\rangle + \frac{1}{R}\langle V_w(u(\cdot,t),w(\cdot,t)), u(\cdot,t)-w(\cdot,t)\rangle + \max(0, -\langle V_w(u(\cdot,t),w(\cdot,t)), w_x(\cdot,t)\rangle)\Big)^2$$

At $u(x,T)=0, w(x,T)=-R(x,t)$, we have: $u(x,T)=0$ and $u\_t(x,T)=0$. Thus, $V(0,-R(x,t))=0$.

$$MSE_u = \big(V(0,-R(x,t))\big)^2 = (V(0,-2))^2$$

At $u(x,T)=0, w(x,T)=-R(x,t)$, V achieves its global maximum.

$$MSE_n = ||\nabla_u V(0,-R(x,t))||_2^2 + ||\nabla_w V(0,-R(x,t))||_2^2$$

Controller:

$$\tilde{\sigma}(t) = \begin{cases} 1, \langle V_w(u(\cdot,t)), w(\cdot,t)\rangle < 0, \\ 0, otherwise \end{cases}$$

Proposed algorithms

HJB value iteration

1. Initiate value network parameter ϕ
2. Run the controller $\tilde{\sigma}(t)$ in the environment for T timesteps and observe samples $\{(s_t, a_t, R_t, s_{\{t+1\}})\}_{t=1}^T$
3. Compute the value network loss as: $J(\phi) = MSE_f + MSE_u + MSE_n$
4. Update $\phi \leftarrow \phi - \alpha\nabla_\phi J(\phi)$
5. Run steps 2-4 for multiple iterations

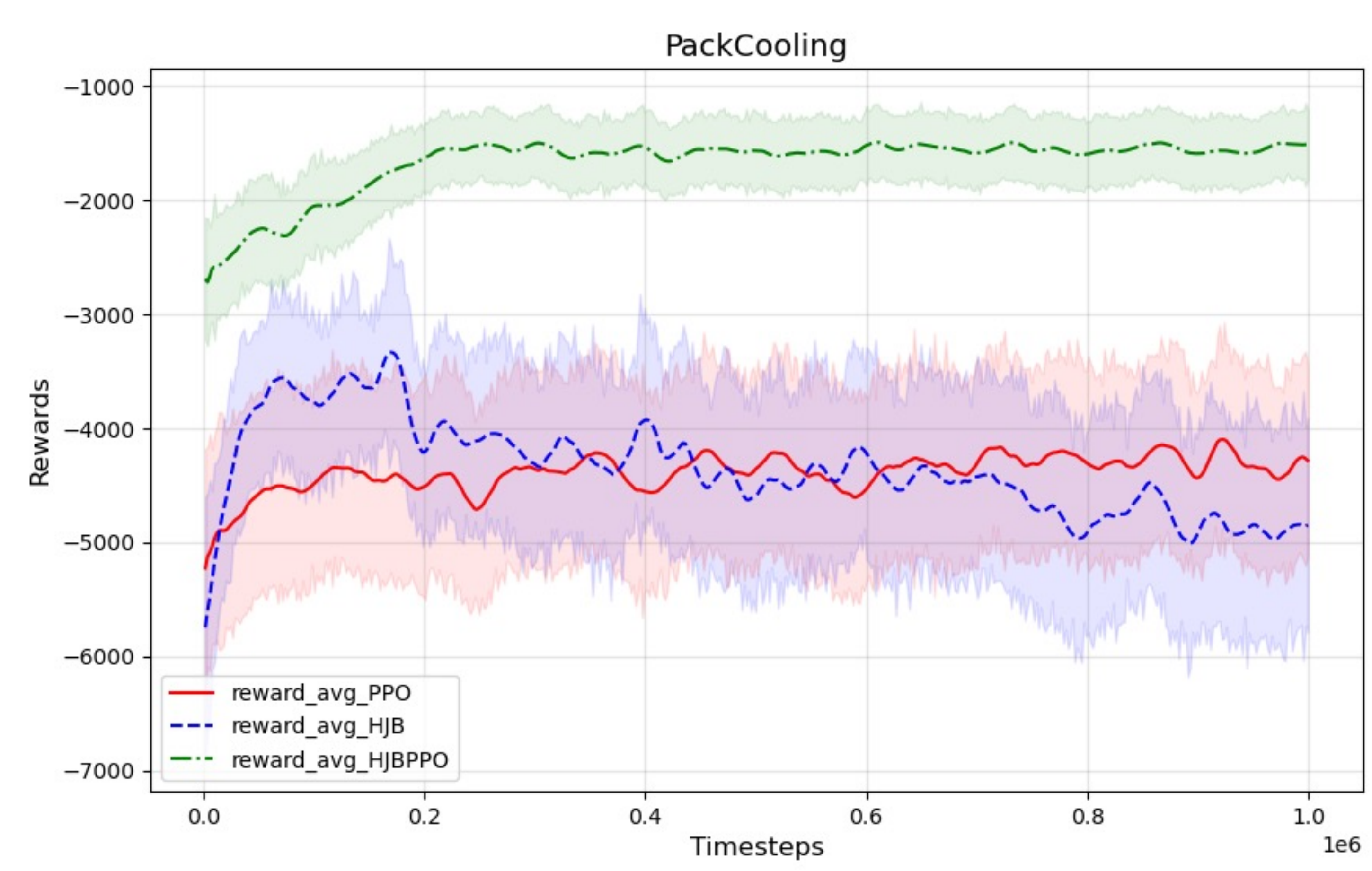
HJBPPO – Action Selection

1. Retrieve state s_t, policy network parameter θ and value network parameter ϕ
2. Sample $i \in \{0,1\}$
3. if $i = 0$ then select the controller $\tilde{\sigma}(t)$
4. else run policy $\pi_\theta(\cdot|s_t)$
5. end

Hamilton Jacobi Bellman Proximal Policy Optimization (HJBPPO)

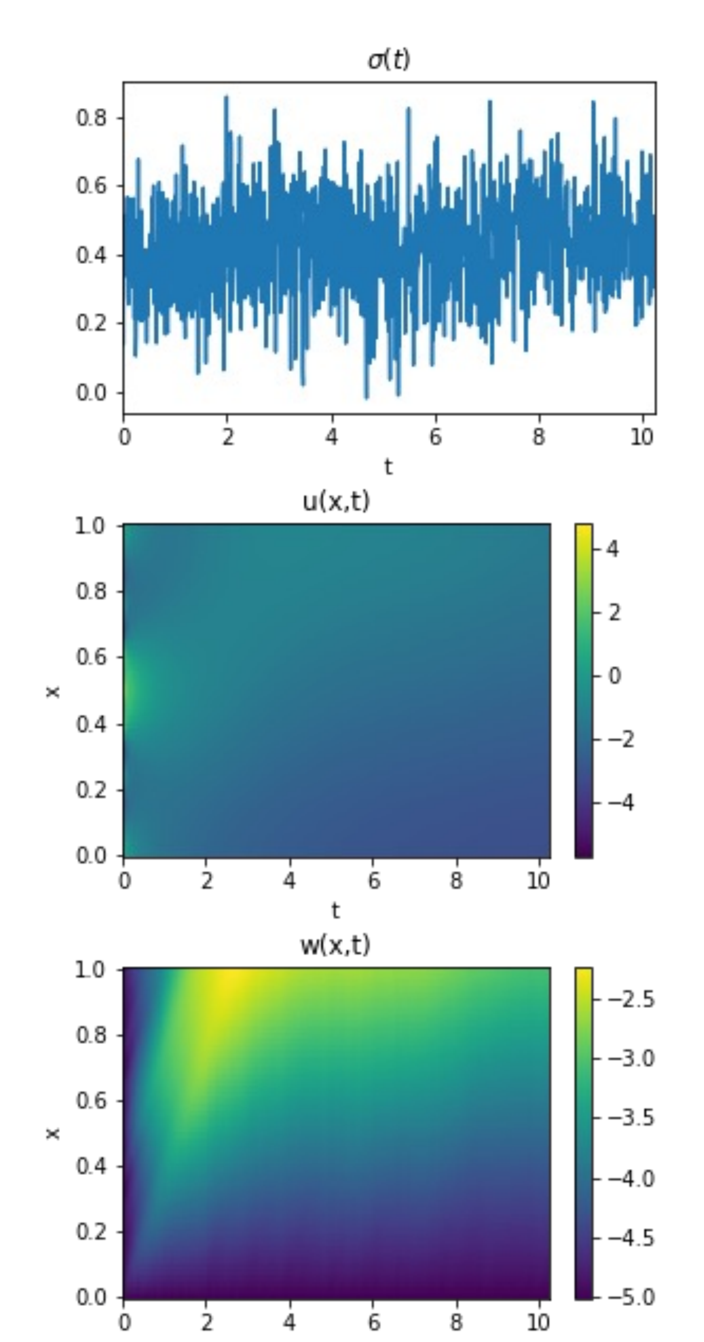
1. Initiate policy network parameter θ and value network parameter ϕ
2. Run action selection as given earlier for T timesteps and observe samples $\{(s_t, a_t, R_t, s_{\{t+1\}})\}_{t=1}^T$
3. Compute the advantage $A_t = \delta t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T-t-1}\delta_{T-1}$ where $\delta = R_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$, γ: Discount factor (≈ 0.99) and λ: Smoothing factor (≈ 0.95)
4. Compute $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$
5. Compute the objective function of the policy network: $L(\theta) = \frac{1}{T}\sum_{t=0}^{T-1}\min[r_t(\theta)A_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t]$ where ε: clipping parameter (≈ 0.2)
6. Update $\theta \leftarrow \theta - \alpha_1\nabla_\phi L(\theta)$
7. Compute the value network loss as: $J(\phi) = MSE_f + MSE_u + MSE_n$
8. Update $\phi \leftarrow \phi - \alpha_2\nabla_\phi J(\phi)$
9. Run steps 2-5 for multiple iterations

Results

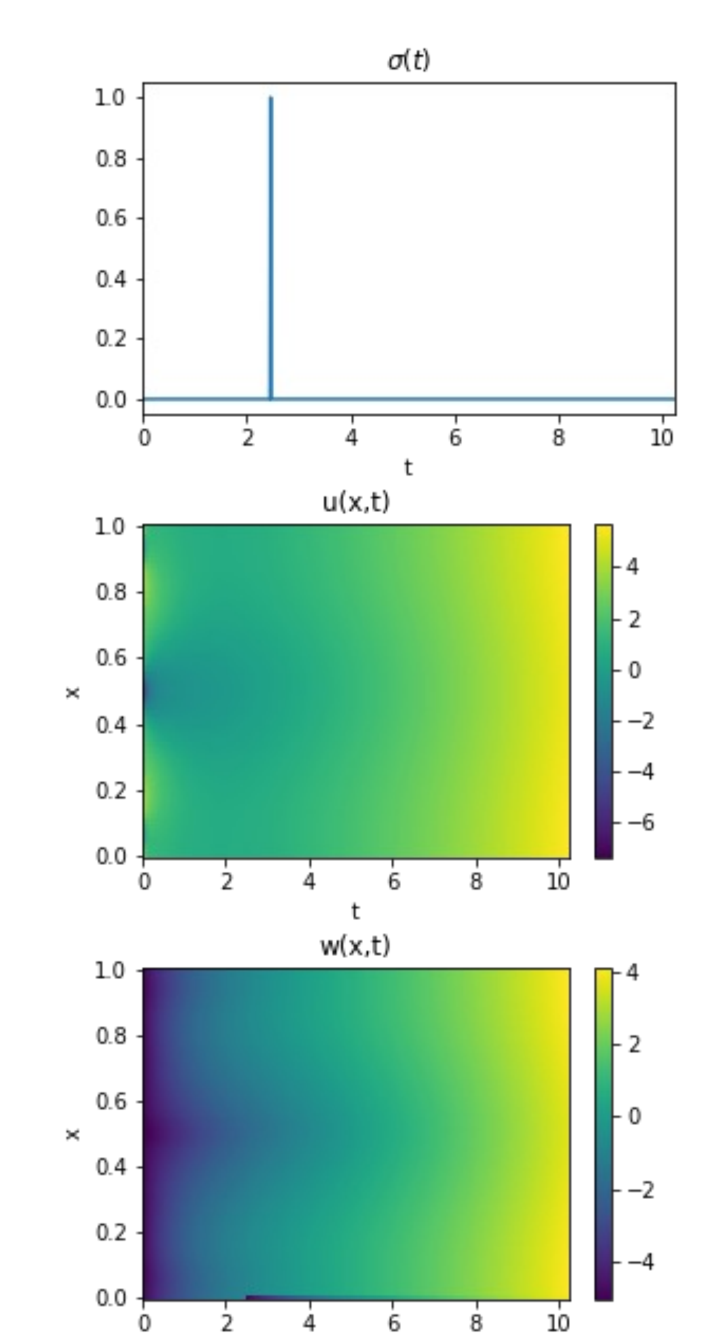


Reward curves of PPO (red), HJB value iteration (blue), and HJBPPO (green) averaged over 5 seeds.
Shaded area indicates 0.2 standard deviations.

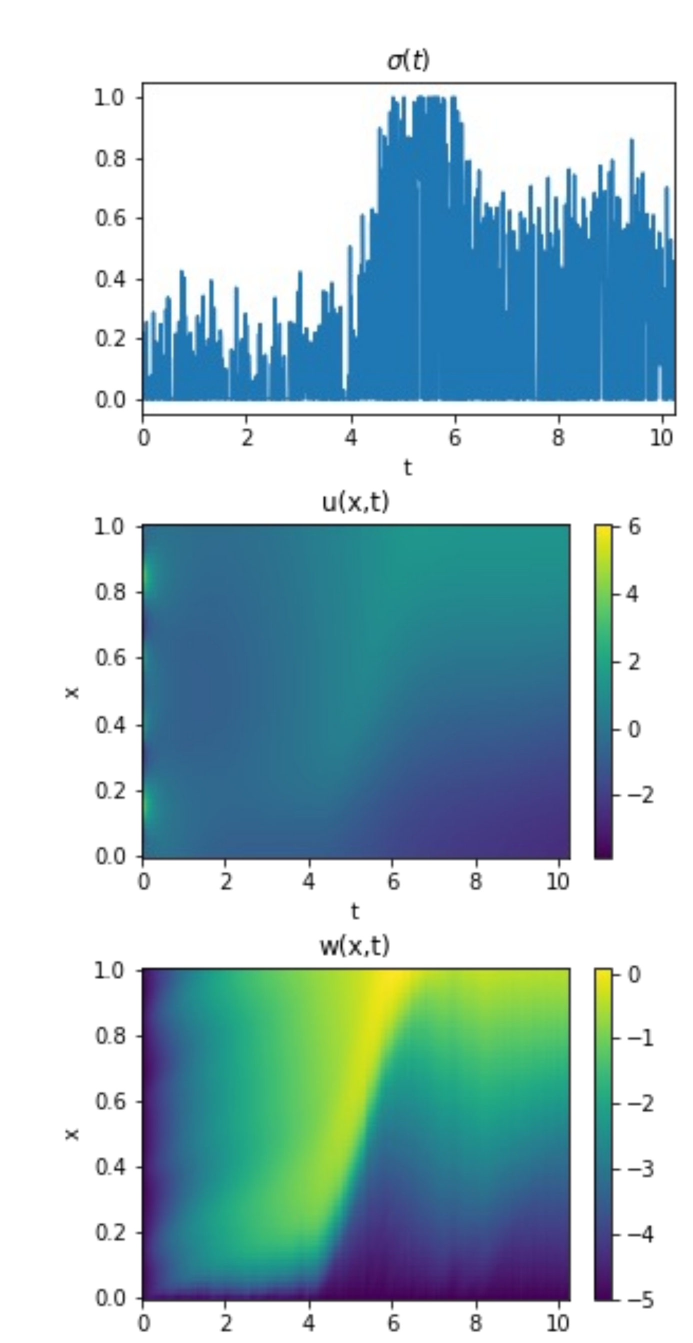
Comparison of Trajectories



Trajectory of PPO. Cummulative reward: -3970.02



Trajectory of HJB value iteration. Cummulative reward: -7294.51



Trajectory of HJBPPO. Cummulative reward: -881.55