Sample Efficient Deep Reinforcement Learning for Optimal Control in Chemical Industrial Processes

Nayoung Ahn
18.10.2023, Aachen
Masters thesis presentation

Examiners: Prof. Alexander Mitsos, Ph.D.
Prof. Dr.-Ing. Adel Mhamdi

Supervisors: Daniel Mayfrank, M.Sc.
Jan Christoph Schulze, M.Sc.

Aachener Verfahrenstechnik | RWTH AACHEN UNIVERSITY

Why do we need an optimal control in chemical industrial processes?

Rise in renewable energy

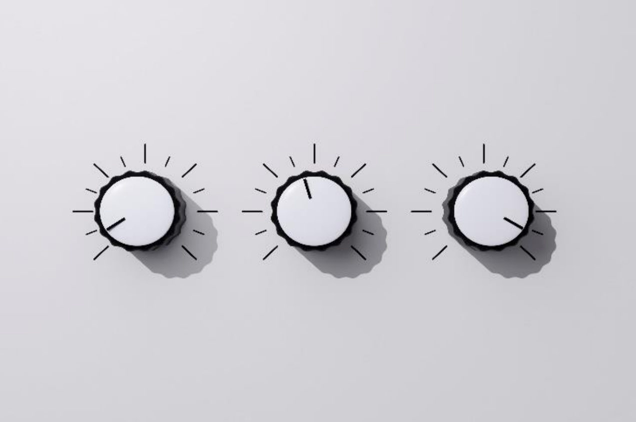
- Volatile
- Fluctuating prices

Steady energy demand

- High potential for saving with demand response

Hard to produce flexibly

- Non-linear dynamics
- Physical constraints

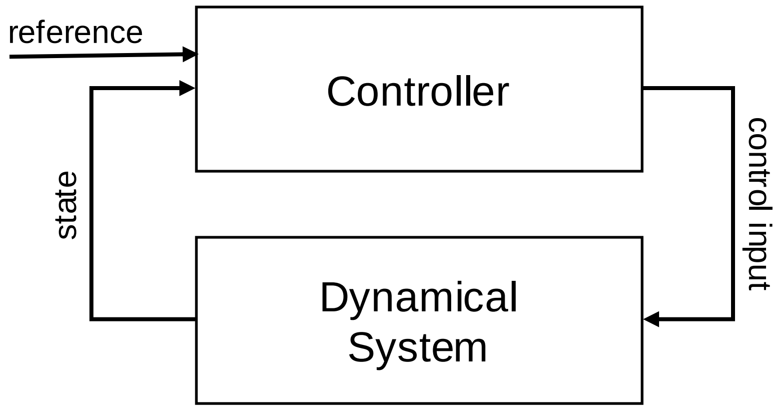


Need an advanced control method that saves electricity costs + adheres to physical constraints

[1] Leinauer et int., Weibelzahl, Energy Policy, 2022.

Why do we turn to deep reinforcement learning approach? – RL as an optimal control method

Model Predictive Control (MPC)



Reinforcement Learning (RL)

Dynamical system	**Dynamics**	Environment
Controller	**Decision unit**	Agent via policy
State, Control input	**Variables**	State, Action
Minimize cost	**Goal**	**Maximize reward**
Model based	**Principle**	**Learn from interaction** with or without model

[1] Brunton et int., Kutz, Cambridge University Press, 2019.

Why do we turn to deep reinforcement learning approach? – concise evolution of approaches

Industrial Demand Response Problem	• Complex dynamics
	  • Non-linear
	  • High dimensional
		• Physical constraints
		• Continuous space
			• Long-term stable
			• High quality yield
			• Economical

Model Predictive Control → **Reinforcement Learning** → **Differential Simulation + RL**

Model Predictive Control

+ sample efficient

– require a system model
– computationally expensive (online)

Reinforcement Learning

+ computationally less expensive
+ system model not necessary

– need vast training data

Differential Simulation + RL

+ need less data
+ promise of better terminal performance

– local minimum
– gradient problems

Need an advanced control method with reduced computational cost + quality control solutions (fast & sample efficient)

[1] Leinauer et int., Weibelzahl, Energy Policy, 2022.
[2] Brunton et int., Kutz, Cambridge University Press, 2019.
[3] Xu et int., Macklin, arXiv preprint, 2022.

What is sample efficiency and why do we need it?

Industrial Demand Response Problem	• Complex dynamics • Non-linear • High dimensional	• Physical constraints • Continuous space	• Long-term stable • High quality yield • Economical

Sample efficiency

- the amount of data required for a learning algorithm to achieve a target performance standard [2]
- In RL, the number of agent-environment interactions essential for deriving an effective policy

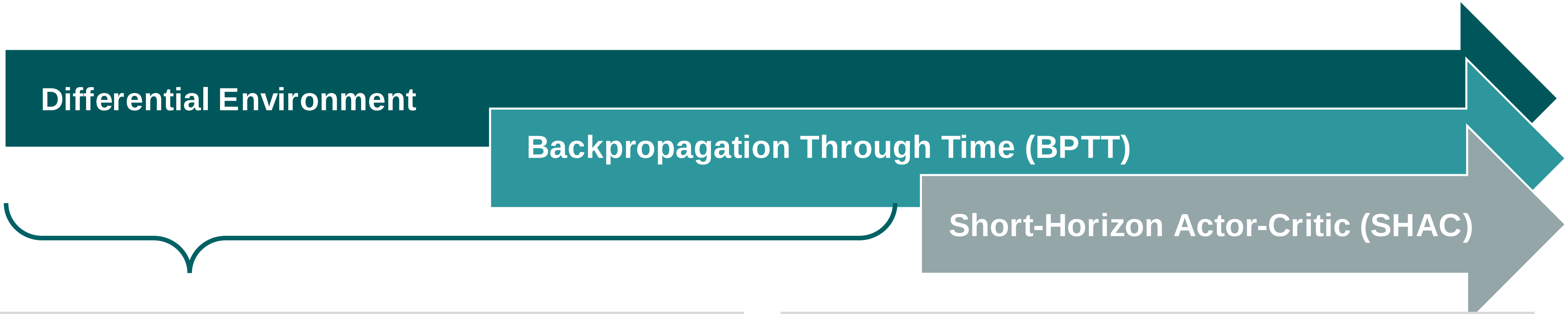
Rephrased objective:

- **Efficient learning** of an effective policy that yields high cost savings and min. constraint violations.
- Trained model performance should be consistent for long-term task horizon, e.g. 8750hrs (=1 yr.)

[1] Leinauer et int., Weibelzahl, Energy Policy, 2022.
[2] Botvinick et int., Hassabis, Trends in Cognitive Sciences, 2019.

What is my proposed approach? – progression of implementation

Objective: **Efficient learning** of an effective policy via gradient-based optimization



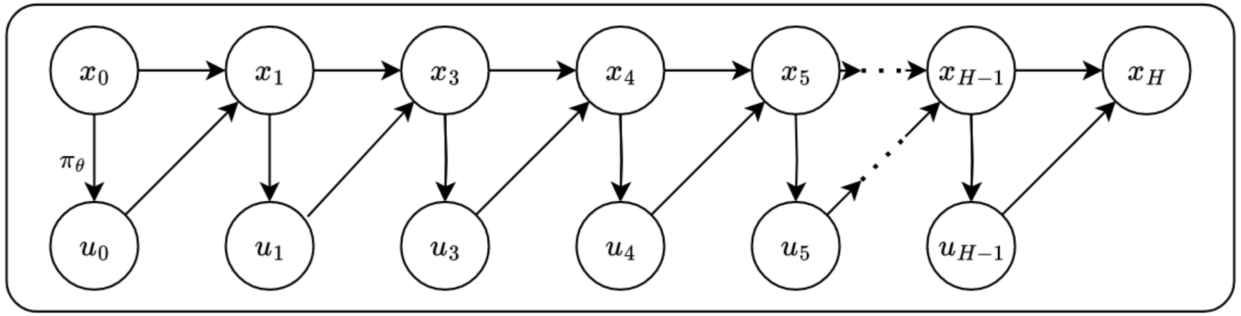
Problem:
- stuck in local minimum
- vanishing, exploding gradients

Solution:
- use a critic network
- shorten backpropagation path via short learning windows

[1] Xu et int., Macklin, arXiv preprint, 2022.

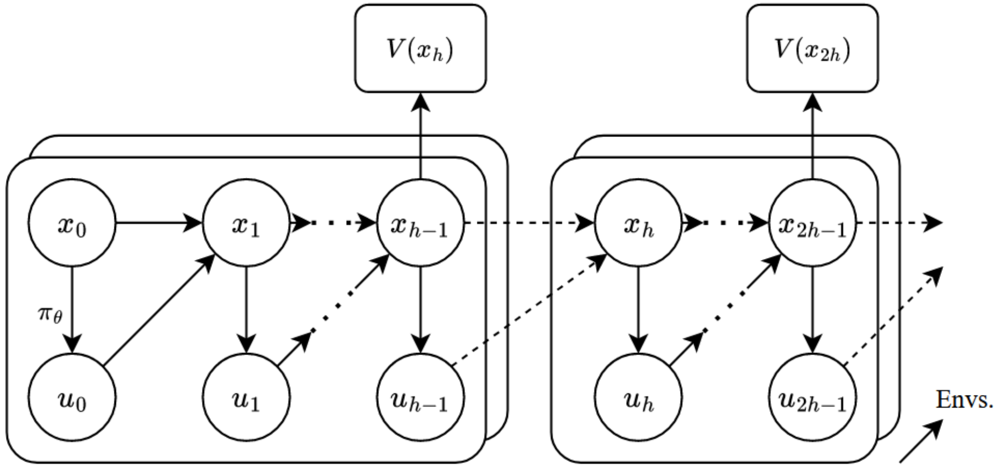
How are the problems with BPTT solved with SHAC? – computation graph comparison

BPTT



Learning episode of horizon length H

SHAC



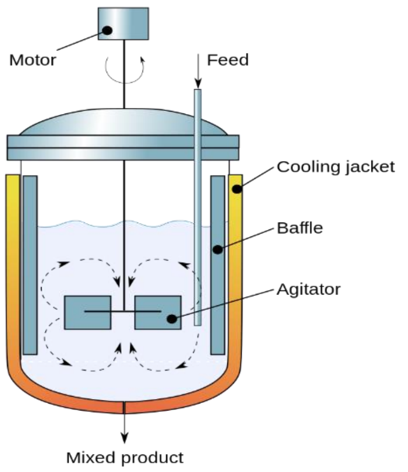
Learning episode of horizon length h **Learning episode of horizon length h**

- Long horizon H → short horizon of length h
- Results in smooth surrogate reward over N parallel environments

- Solid arrows denote gradient-preserving computation
- Dashed arrows, where gradients are cut off

[1] Xu et int., Macklin, arXiv preprint, 2022.

Case Study: Continuous-Stirred Tank Reactor (CSTR)



Cross-sectional diagram of a CSTR
By Daniele Pugliesi - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6915706

	Symbol	Value
Volume	V	20
Reaction constant	k	$300h^{-1}$
Activation energy	N	5
Feed temperature	T_f	0.3947
Heat transfer coefficient	α_c	1.05×10^{-4}
Coolant temperature	T_c	0.3816

Dynamic Equation:

$$\dot{c}(t) = [1 - c(t)] \cdot \frac{\rho(t)}{V} - c(t) \cdot k \cdot e^{\left(-\frac{N}{T(t)}\right)}$$

$$\dot{T}(t) = [T_f - T(t)] \cdot \frac{\rho(t)}{V} + c(t) \cdot k \cdot e^{\left(-\frac{N}{T(t)}\right)} - F_c(t) \cdot \alpha_c[T(t) - T_c]$$

		Symbol	Steady State	Lower lim.	Upper lim.
System states	Concentration	c	0.1367	$0.9 \times c_{ss}$	$1.1 \times c_{ss}$
	Temperature	T	0.7293	0.6	0.8
Control inputs	Material flow rate	ρ	1.0/h	0.8/h	1.2/h
	Coolant flow rate	F_c	390.0/h	0.0/h	700.0/h
	Storage level	l	1	0	24

Observation $\boldsymbol{o}$ = [system states $\boldsymbol{x}$, storage level l, relative prices $\boldsymbol{p}_{rel}$]

Price prediction $\boldsymbol{p}$ ➡ $\boldsymbol{p}_{rel} = \boldsymbol{p} - \boldsymbol{EMA}(\boldsymbol{p})$

ODE solver:

odeint

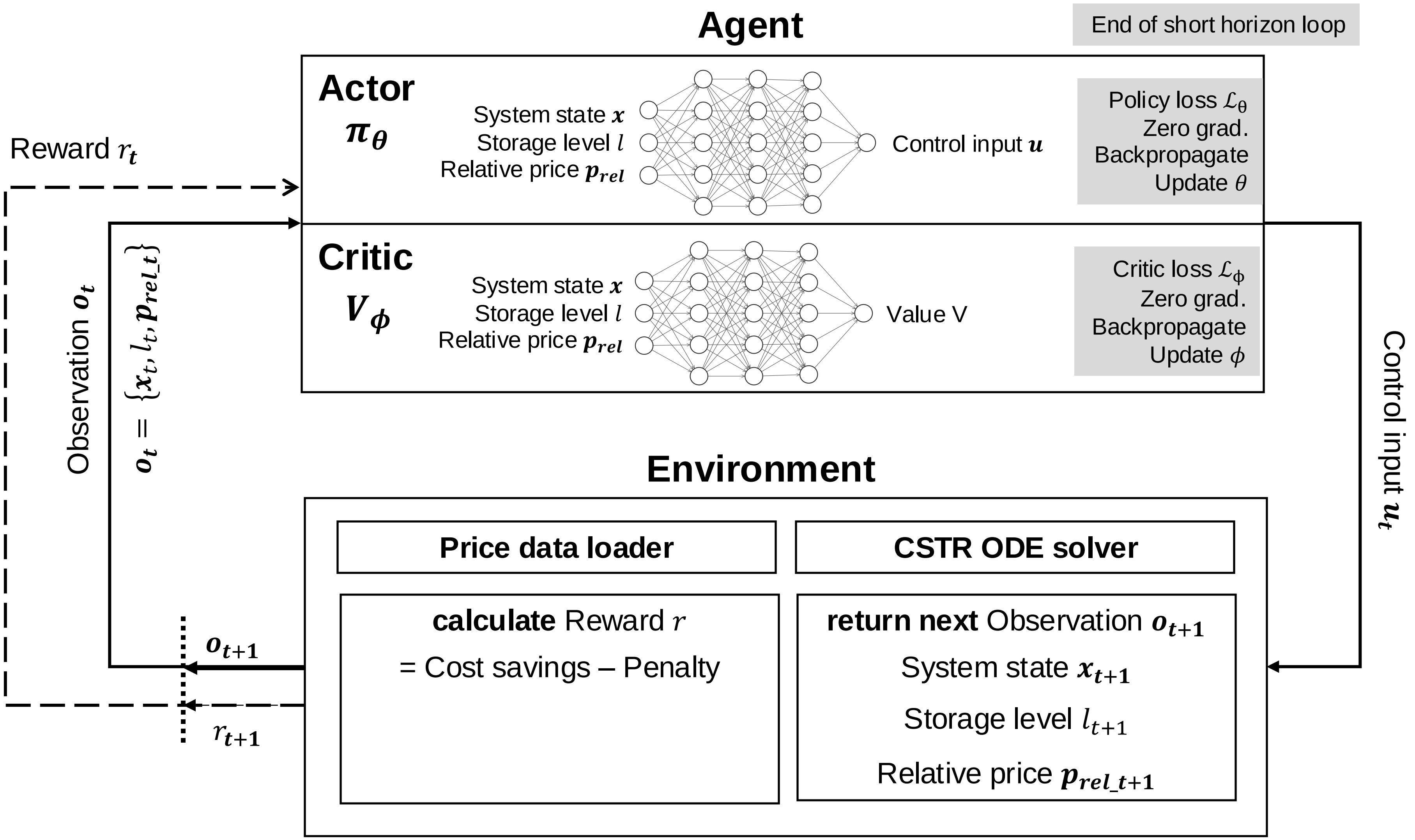
↓

torchdiffeq

Integration method: rk4

[1] Flores et int., Grossmann, Industrial & Engineering Chemistry Research, 2006.
[2] Chen et int., Duvenaud, Advances in Neural Information Processing Systems, 2018.
[3] Virtanen et int., Mulbregt, Nature Methods, 2020.

Final implementation: Short-Horizon Actor-Critic (SHAC)



Settings for a comparative analysis for terminal performance

- Aim to **maximize cost savings**, while **minimizing constraint violations** with **minimal training episodes**.
- All models have been empirically tuned to train smoothly at least until 20,000 episodes.
- Trained with same price profiles, similar policy network architecture, and used Adam optimizer
- Tested on same price profile of 1 year task horizon.

	BPTT	SHAC	PPO
No. of training episodes	10,000	10,000	10,000
Optimizer	Adam	Adam	Adam
Training horizon length	72	32	72
No. of parallel environments	1	6	1
Steps between updates to agent	72	192	2048
Minibatch size	N/A	N/A	64

[1] Mayfrank et int., Dahmen, arXiv preprint, 2023.

Comparison of terminal performance after 10,000 training episodes

Performance on a price horizon of 1 year

	BPTT	SHAC	PPO
Relative cost savings [%]	19.40	14.07	8.7
Penalties occurrence [%]	99.16	15.66	0.00
Avg. contraint violation size	12.08	0.001	0.000
Avg. storage level	-287.35	1.47	0.48

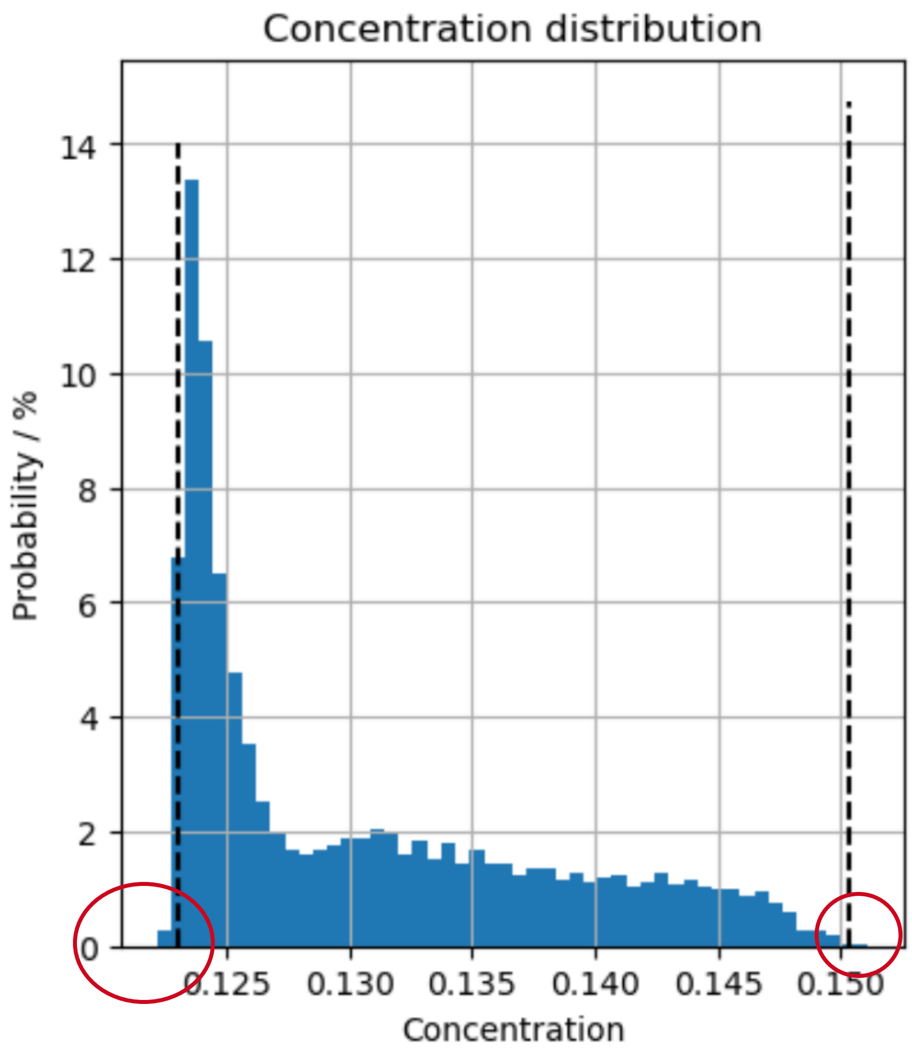
Performance on a price horizon of 6 mths on new data

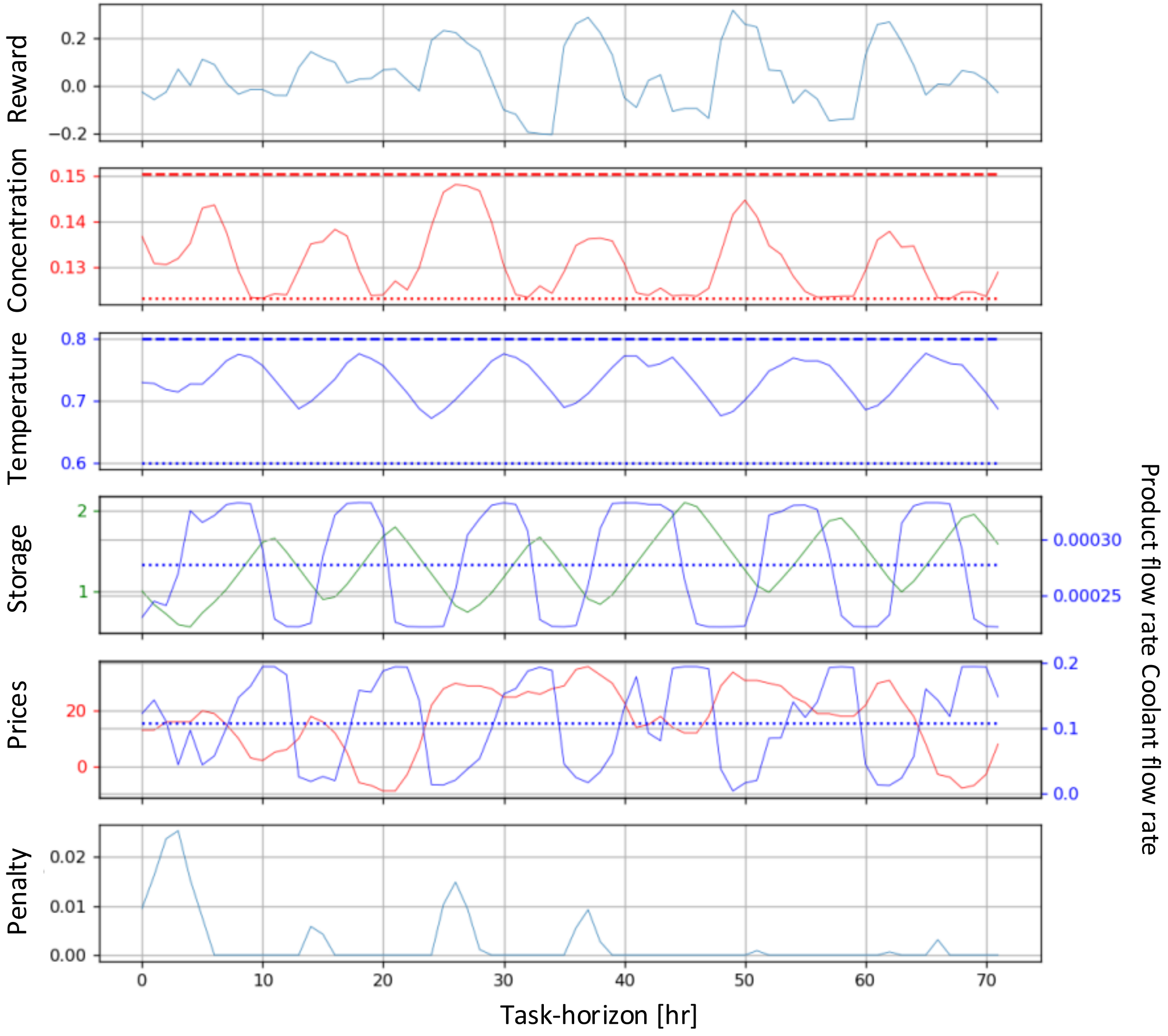
	BPTT	SHAC	PPO
Relative cost savings [%]	17.47	12.02	6.00
Penalties occurrence [%]	99.93	19.95	0.22
Avg. contraint violation size	5.968	0.001	0.000
Avg. storage level	-140.52	1.39	5.32

- SHAC shows good balance between cost savings and adhering to constraints.
- SHAC‘s performance is consistent when tested on new data
  - March 26,2018 – September 30, 2018

[1] Mayfrank et int., Dahmen, arXiv preprint, 2023.

Control behavior for a test task-horizon of 72 hrs





What has been my contribution to this topic?

What was given?

- Problem statement
- SHAC paper
- A PPO implementation
- A CSTR ODE solver using scipy

How did I go beyond my goal?

- Design comparison experiments
- Design performance metrics for control performance
- Suggested to learn from relative price movement
- Set up a clean code base
- Created realtime plots and test notebooks for analysis

What planned goals did I achieve?

- Implemented a fully differentiable environment
  - CSTR ODE solver using torchdiffeq
  - Differentiable reward
  - Differentiable penalty
- Implemented BPTT
- Implemented SHAC

What would be a suitable follow-up?

- Training:
  - Rigorous hyperparameter search
  - Train for more than 100,000 episodes
  - Track wall clock time
  - Run in parallel with access to GPU
- Testing:
  - On new data with more volume from other region
  - Compare best trained model for control performance, sample efficiency and wall clock time

Conclusion

- Implemented a fully differentiable environment
- Implemented state of the art algorithm SHAC
- Achieved 14% cost efficiency over one-year task-horizon
- While keeping constraint violation to the minimum

However, challenges remain toward true sample efficiency.

Rigorous hyperparameter tuning, experimentation and validation are recommended.

Vielen Dank für Ihre Aufmerksamkeit