Applying Reinforcement Learning to Process Control

Debangsu Bhattacharyya

Department of Chemical and Biomedical Engineering, West Virginia University, Morgantown, WV

Adjunct Professor, Department of Chemical Engineering, Carnegie Mellon University, Pittsburg, PA

10/23/22

ESI Seminar

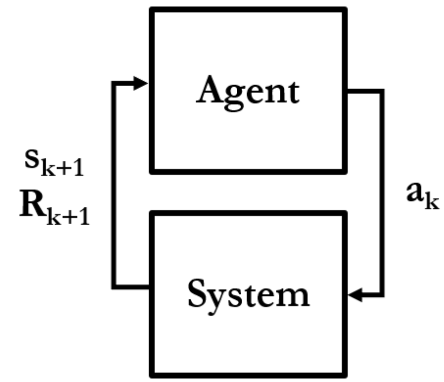
Carnegie Mellon University

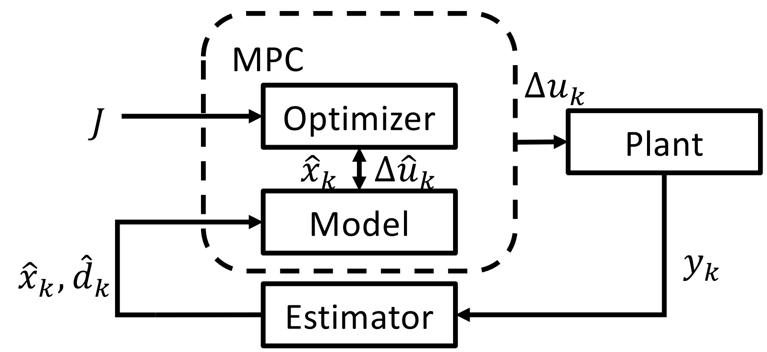
West Virginia University

BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Motivation and Background

- Reinforcement learning (RL) is a machine learning method that learns from active sampling of system performance
- Integration of RL with a process controller such as PID or MPC can exploit strengths of both and addresses some of the weaknesses
- RL can also be used by itself at the supervisory or higher layer controller





RL Basics

- Learning based on a value function and/or a control policy
  - Algorithms with a fixed policy – focused on learning a value function given the fixed policy (e.g., Q-Learning, SARSA)
  - Some algorithms where the policy is learned with a value function – actor-critic methods; parameterized policy and value function used for control an updates
  - General goal is to maximize expected sum of rewards:

$$Q^{\pi}(s,a) = E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\}$$

RL Basics: State-action-reward-state-action (SARSA)

$$Q^{\pi}(s,a) = E\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right\}$$

- Episodic learning for tasks with fixed starting and terminal states
- Continuing learning where these cannot be defined

$$\widehat{\boldsymbol{q}}(s,a) \;=\; w^T q = \sum_{i=1}^{d} w_i q_i$$

$$q_i(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\|c_i - x\|^2}{2\sigma^2}}$$

$$\delta = R + \gamma Q(s_{k+1}, a_{k+1}) - Q(s_k, a_k)$$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\widehat{\boldsymbol{q}}(s_k, a_k)$$

$$a_k = \underset{a}{\operatorname{argmax}}\, \widehat{\boldsymbol{q}}(s,\cdot)$$

Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction, 2nd ed. Bradford, Cambridge, MA, USA.

Tuning Learning and Learning Metrics

- RL has multiple hyperparameters:

$$\alpha \in (0,1], \gamma \in [0,1], \varepsilon \in [0,1]$$

  - α is the learning rate, γ is the discount factor
  - ε controls the rate of exploration vs. exploitation under an ε-Greedy policy
- Goal is to achieve stable learning to an optimal policy
  - Learn Q(s,a) such that action (a) can be selected greedily for all states (s)
- Results show in terms of episode return (G):

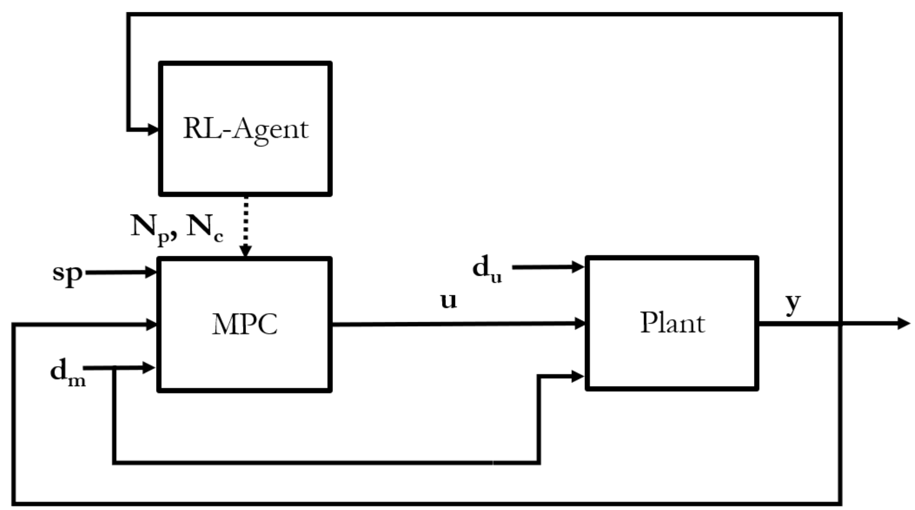
$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

WestVirginiaUniversity
BENJAMIN M. STATLER COLLEGE OF
ENGINEERING AND MINERAL RESOURCES

RL MPC

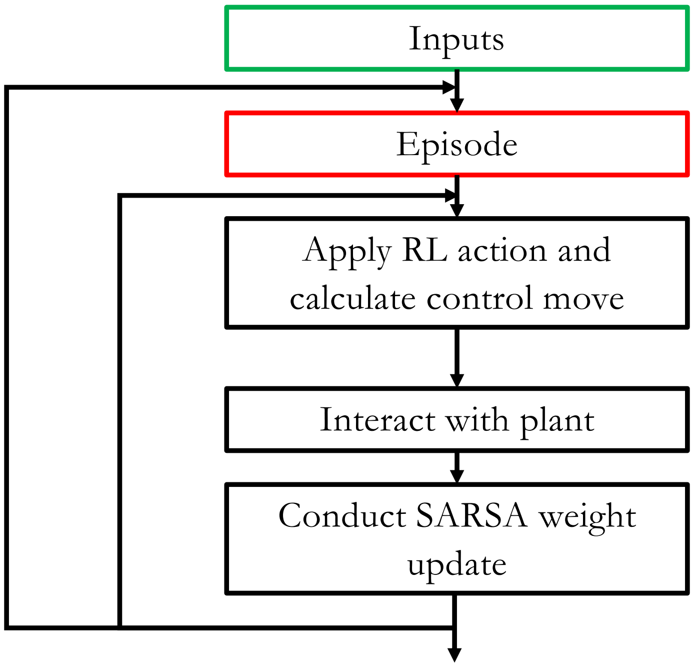
- RL agent is applied to select controller tuning parameters online
  - Underlying MPC controls the plant
- Output feedback to the agent is the RL state and is the subject of the RL reward function:

$$R_k \equiv -\left\| y_{sp} - y_{k+1} \right\|^2$$



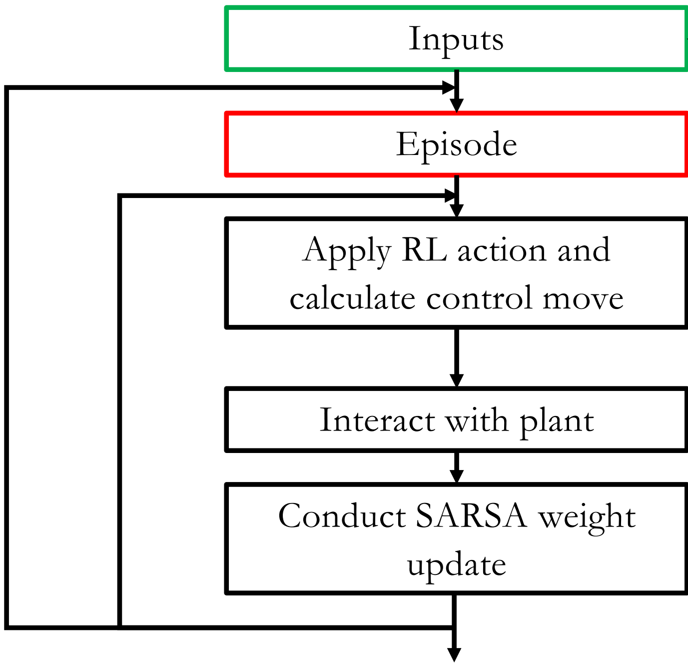
Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm



Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, “Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit,” *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm

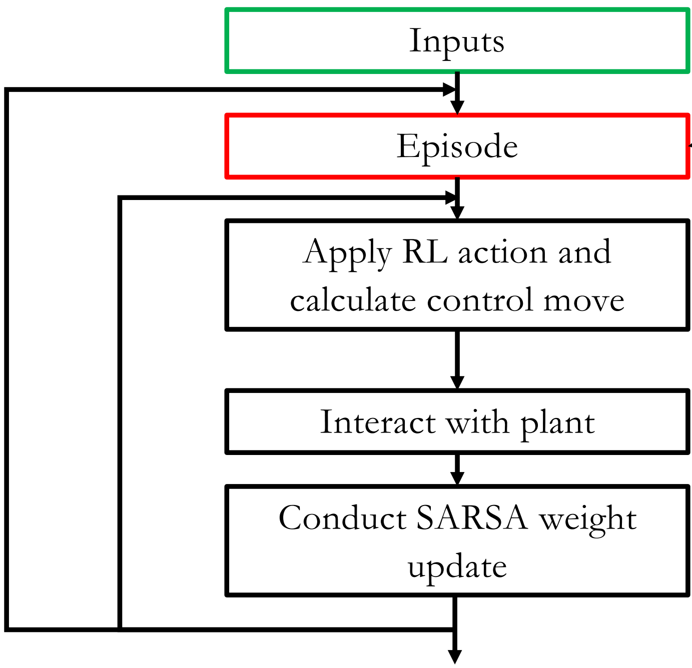


- $\hat{q}(s, a)$ a value function approximator with weights w
- (α, ε) $\in$ (0,1], $\gamma \in$ [0,1)
- Arbitrarily initialize w, the weights of the function approximation
- An MPC to control the plant

WestVirginiaUniversity.
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm



- **Initialize the plant and controller to steady-state (i.e. y = 0, u = 0 in deviation variables)**
- **Selection an initial action (a_0) under the current policy (i.e. ε-Greedy)**

ε-Greedy Search Algorithm

Inputs:

ε ϵ (0,1]

For each evaluation:

Draw P from a uniform distribution

If P > ε

$$a_k = \underset{a}{\operatorname{argmax}}\,\hat{\boldsymbol{q}}(s,\cdot)$$

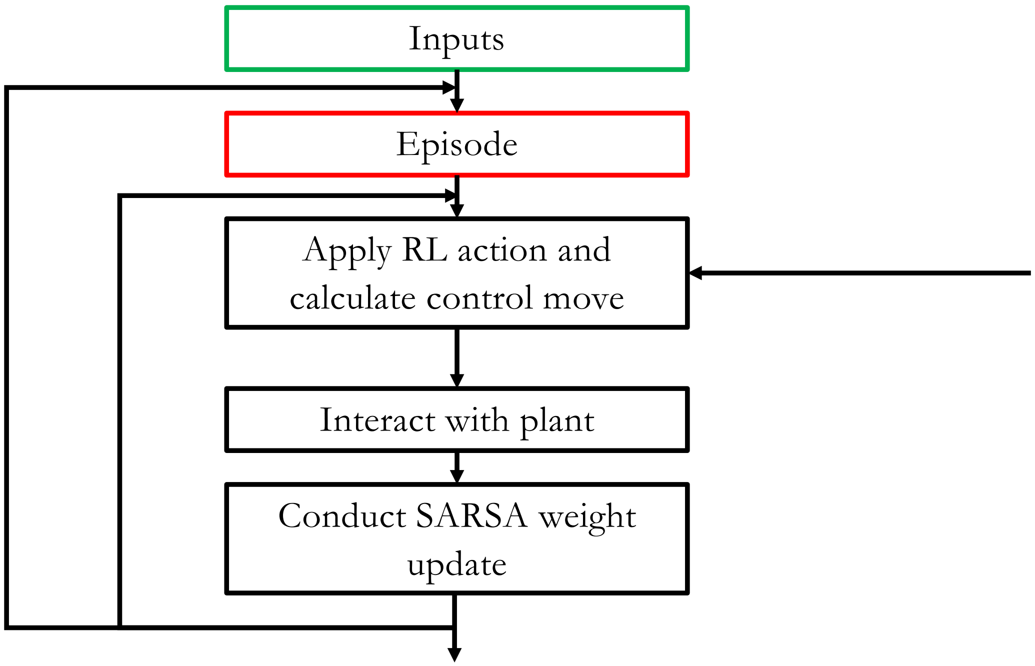
If P < ε

$$select\ a_k = a \in A\ randomly$$

Break ties randomly

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm

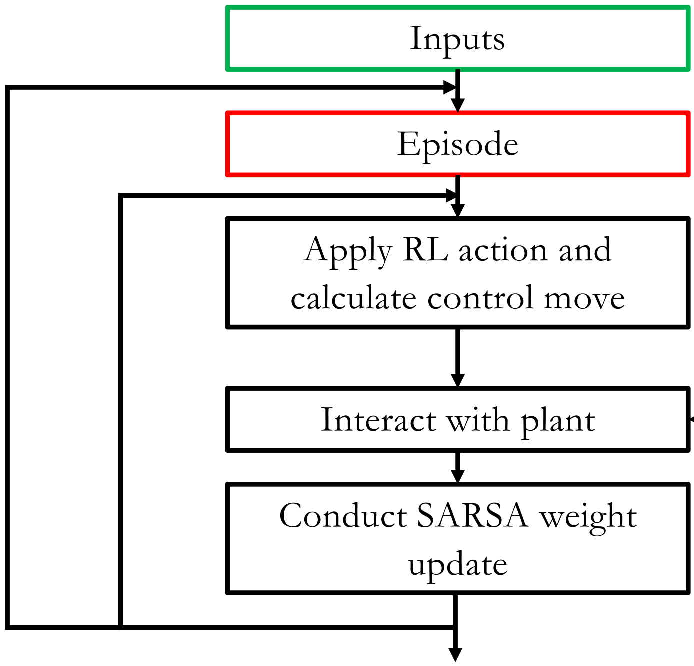


- **Apply** $\boldsymbol{a_k} = \begin{bmatrix} \boldsymbol{N_p} \\ \boldsymbol{N_c} \end{bmatrix}_{\boldsymbol{k}}$ **to the MPC**
- **Calculate:**

$$\boldsymbol{u_k} = \boldsymbol{u^{*} \, from \; the \; MPC}$$

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

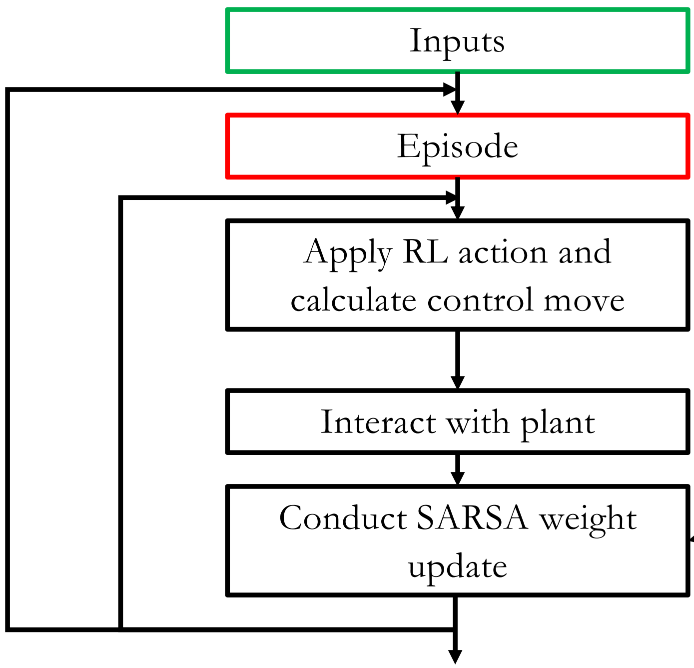
RL-MPC Algorithm



- **Apply u_k to the plant and get y_{k+1}**
- **Calculate $R_k = R(y_{k+1})$**

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm



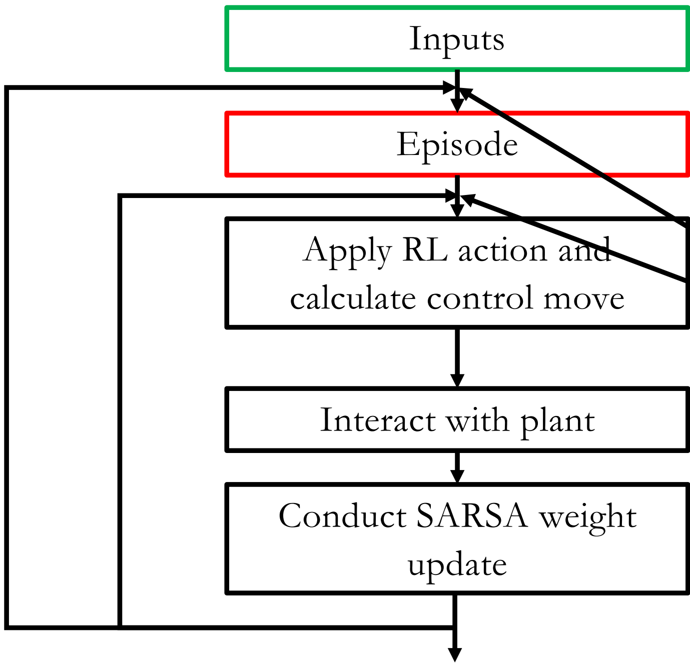
State-action-reward-state-action (SARSA)

- **Select a_{k+1} under the current policy (i.e. ε-Greedy)**
- **Calculate $\delta \leftarrow R_k + \gamma \hat{q}(y_{k+1}, a_{k+1}) - \hat{q}(y_k, a_k)$**
- **Update $w \leftarrow w + \alpha \delta \nabla \hat{q}(y_k, a_k)$**

WestVirginiaUniversity.
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm

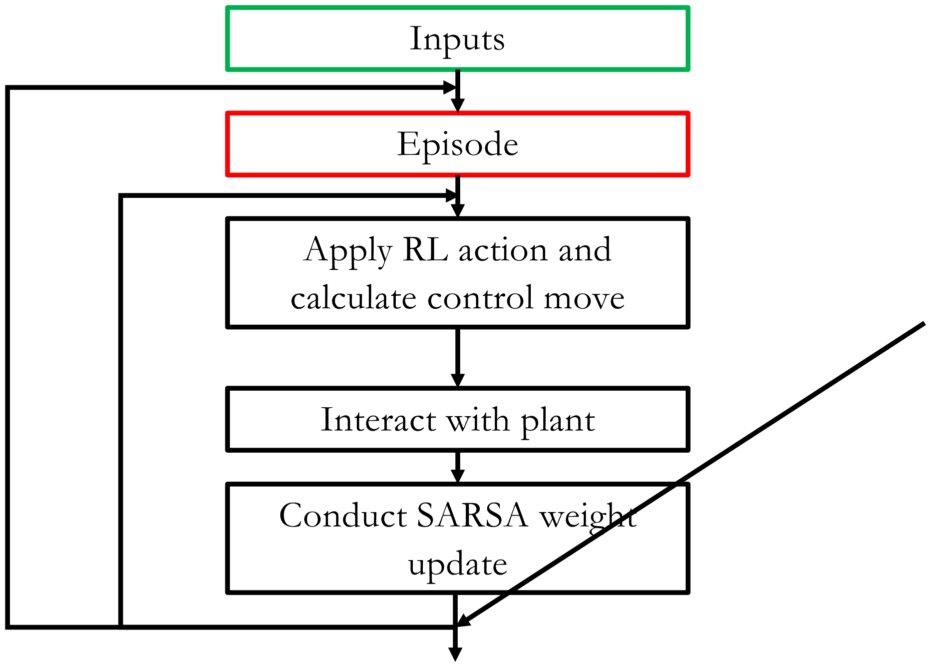


- **For each episode continue until the terminal state (or time) is reached**
- **Continue until a maximum number of episodes is reached**

WestVirginiaUniversity.
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

RL-MPC Algorithm



- **Output is a weight vector which can be used for action selection, under the same policy, in an online setting**
- **Continuous learning can be carried out when implemented on true plant**

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

Offline RL-MPC Algorithm

Inputs:

$\hat{q}(s, a)$ a value function approximator with weights **w**

$\alpha \epsilon (0,1]$, $\varepsilon \epsilon [0,1]$, $\gamma \epsilon [0,1]$

Arbitrarily initialize **w**, the weights of the function approximation

An MPC to control the plant

For each episode:

Initialize the plant and controller to steady-state (i.e. y = 0, u = 0 in deviation variables)

Selection an initial action (a_0) under the current policy (i.e. ε-Greedy)

For each timestep (k) of each episode:

1. Apply $a_k = \begin{bmatrix} N_p \\ N_c \end{bmatrix}_k$ to the MPC
2. Calculate:

   $u_k = u^* from\ the\ MPC$
3. Apply u_k to the plant and get y_{k+1}
4. Calculate $R_k = R(y_{k+1})$
5. Select a_{k+1} under the current policy (i.e. ε-Greedy)
6. Calculate $\delta \leftarrow R_k + \gamma \hat{q}(y_{k+1}, a_{k+1}) - \hat{q}(y_k, a_k)$
7. Update $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\boldsymbol{\nabla}\hat{\boldsymbol{q}}(\boldsymbol{y_k}, \boldsymbol{a_k})$

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

West Virginia University
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Online RL-MPC Algorithm

Inputs:

$\hat{q}$(s,a) that is differentiable, consistent with offline learning

$\alpha \in (0,1]$, $\varepsilon \in [0,1]$, $\beta \in (0,1]$

$\mathbf{w}$ from offline learning, $\overline{\mathrm{R}} = 0$

An MPC to control the plant, consistent with offline learning

For each sampling time (k), continuing while online:

1. Apply $a_k = \begin{bmatrix} N_p \\ N_c \end{bmatrix}_k$ to the MPC
2. Calculate:

   $u_k = u^*$ from the MPC
3. Apply u_k to the plant and get y_{k+1}
4. Calculate $R_k = R(y_{k+1})$
5. Select a_{k+1} under the current policy (i.e. ε-Greedy)
6. Calculate $\delta \leftarrow R - \overline{R} + \hat{q}(y_{k+1}, a_{k+1}, w) - \hat{q}(y_k, a_k, w)$
7. Update $\overline{R} \leftarrow \overline{R} + \beta\delta$
8. Update $\mathbf{w} \leftarrow \mathbf{w} + \alpha\delta\nabla\hat{\mathbf{q}}(\mathbf{y_k}, \mathbf{a_k}, \mathbf{w})$

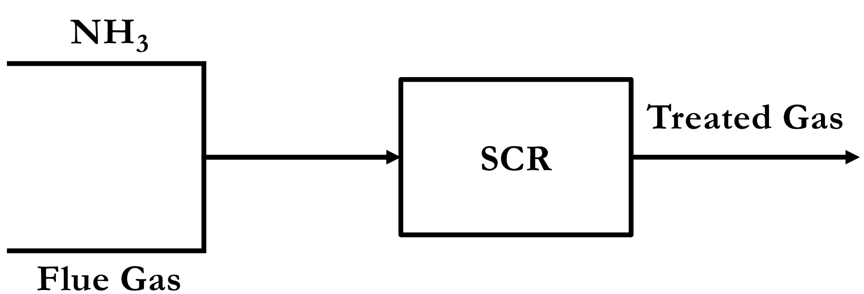
Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

Motivating Example – SCR Control

- Control of an industrial Selective Catalytic Reduction (SCR) reactor for NOx emission reduction in a coal-fired power plant is taken as an example for this work
  - SCR model is a 1D heterogeneous plug flow reactor model with detailed kinetics

$$4\ NH_3 + 4\ NO + O_2 \rightarrow 4\ N_2 + 6\ H_2O$$

$$4\ NH_3 + 3O_2 \rightarrow 2\ N_2 + 6\ H_2O$$



Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

WestVirginiaUniversity
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

MPC of SCR

- SCR has complex dynamics with many disturbance variables
  - Adsorption/desorption kinetics lead to significant nonlinearity and time delay
  - Some of these variables are measurable, can be modeled
  - Others are not incorporated into control model

Inputs (MVs)	
Inlet Ammonia Flow	u_1
Outputs (CVs)	
Outlet NO_x Concentration	y_1
Outlet Ammonia Concentration	y_2
Modeled Disturbances	
Inlet Flue Gas Flow	d_1
Inlet NO_x Concentration	d_2
Unmodeled Disturbances	
Inlet Ammonia Temperature	d_3
Inlet Flue Gas Temperature	d_4
Inlet Dilution Air Flow	d_5

MPC of SCR for Disturbance Rejection – MPC-1

- Initial MPC formulation uses a linear model and a disturbance rejection objective
  - Servo control not regularly needed for SCR operation
- One-step ARX model identified using least-squares estimate

$$x_{k+1} = Ax_k + Bu_k$$

$$y_k = Cx_k, C = I$$

$$\Phi = (\Psi^T\Psi)^{-1}\Psi^T Y$$

$$\min_u J = y_{N_p}^T H y_{N_p} + \sum_{k=0}^{N_p-1} y_k^T Q y_k + \sum_{k=0}^{N_c} u_k^T D u_k$$

$$s.t. x_{k+1} = Ax_k + Bu_k$$

$$y_k = Cx_k$$

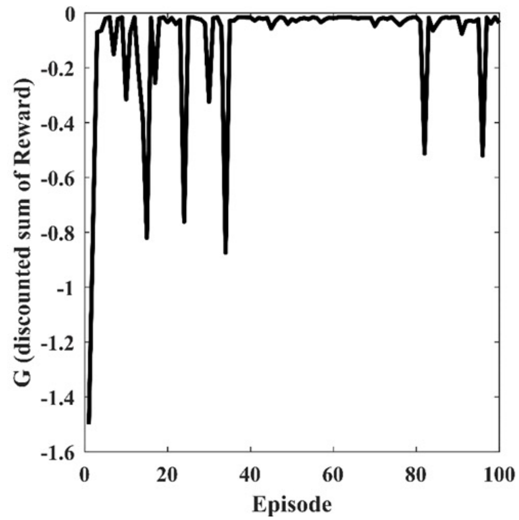
$$u_{i,lb} \leq u_{i,k} \leq u_{i,ub}$$

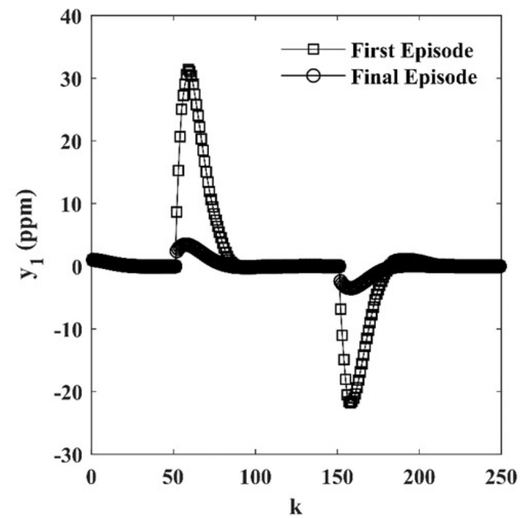
Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

WestVirginiaUniversity.
BENJAMIN M. STATLER COLLEGE OF
ENGINEERING AND MINERAL RESOURCES

Learning Results

- To standardize learning results, the algorithm has been applied for 20 randomized trials
- In early episodes of learning, performance is poor both because exploration is intentionally high and because the agent has little knowledge of the system
- High final value shows learning

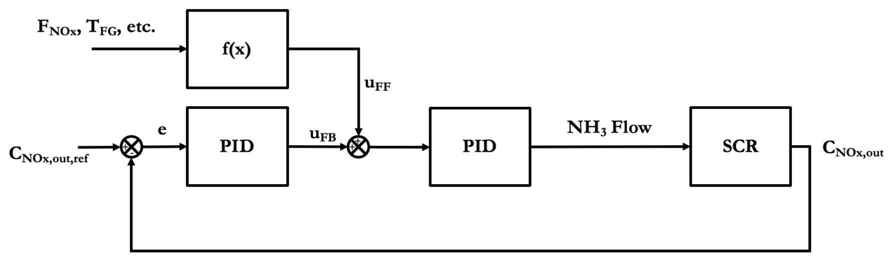




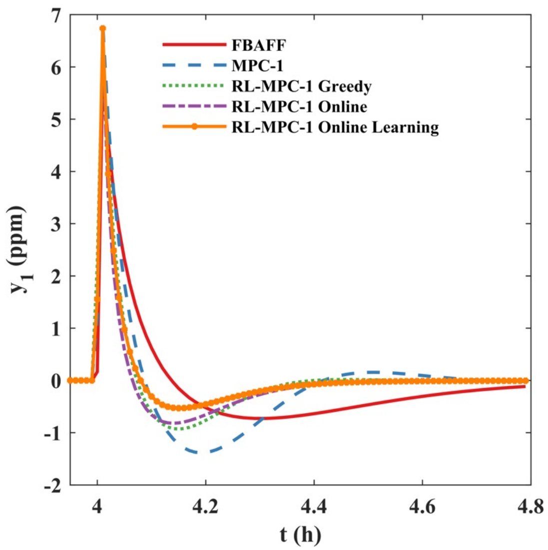
Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

Online Control Results

- Studies for disturbance rejection
  - Unmodeled disturbances in flue gas temperature (d_4)
  - Load following: simultaneous variation of all disturbance variables following industrial profile
- Results compared with the industry-standard feedback augmented feedforward (FBAFF) controller



Online Control Results for (unmodeled) Disturbance in Flue Gas Temperature

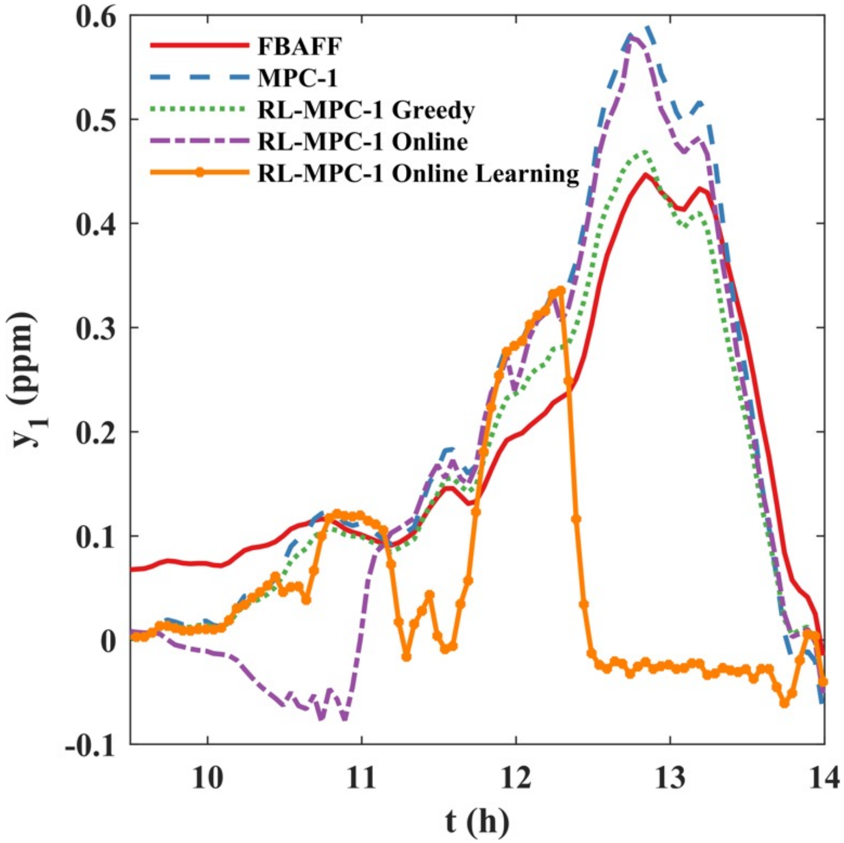


Controller	ISE	Ratio to FBAFF	IAE	Ratio to FBAFF
FBAFF	622	--	300	--
MPC-1	685	1.10	276	0.92
RL-MPC-1 Greedy	462	0.74	255	0.85
RL-MPC-1 Online	453	0.73	250	0.83
RL-MPC-1 Online Learning	384	0.62	241	0.80

Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

West Virginia University
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Load Following (Flue gas flowrate, temperature and inlet NO_X concentration all changing) Control Results

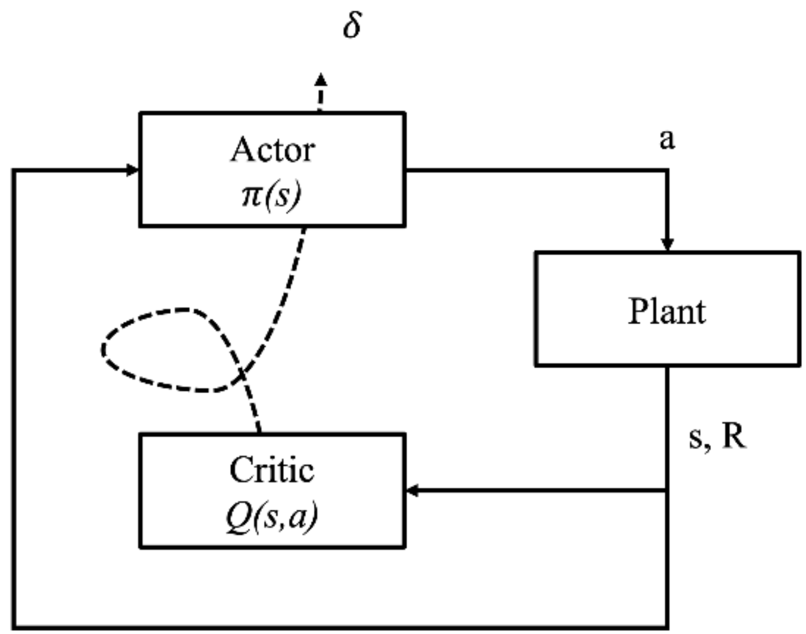


Publication: E. Hedrick, K. Hedrick, D. Bhattacharyya, S. E. Zitney, and B. Omell, "Reinforcement learning for online adaptation of model predictive controllers: Application to a selective catalytic reduction unit," *Comput. Chem. Eng.*, vol. 160, p. 107727, 2022, doi: 10.1016/j.compchemeng.2022.107727.

Offset-Free Actor-Critic Control Development

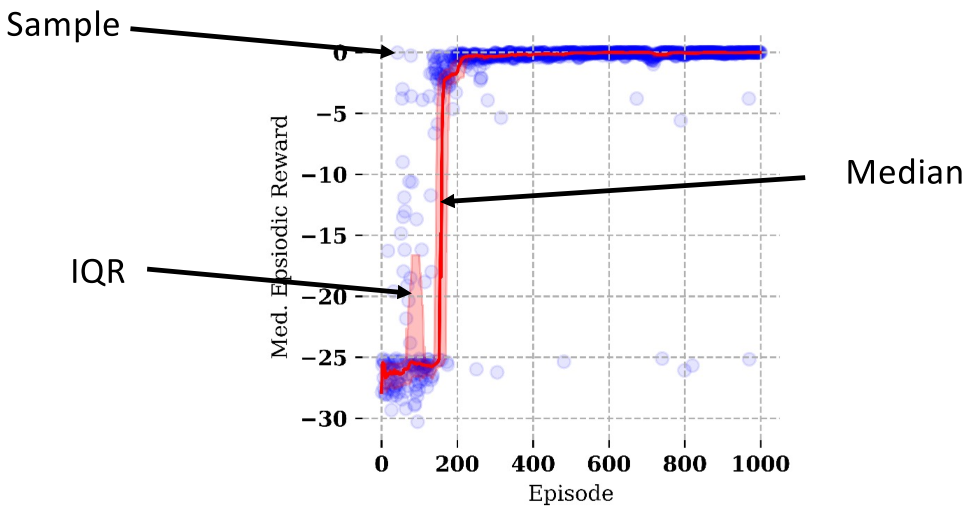
Actor-Critic Approaches

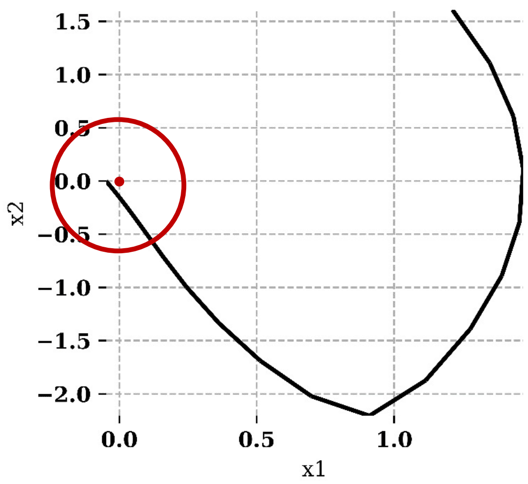
- Use of a parameterized value function, Q, and a parameterized policy, π, for control
- Critic weight update is the same as before
- Policy updated by the current critic
- Actor and critic, most commonly, are deep neural networks



Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: An Introduction, 2nd ed. Bradford, Cambridge, MA, USA.

Control with Offset (Naïve Deep Deterministic Policy Gradient (DDPG))





Learning with DDPG

- Target value-function network Q and target policy network μ are parameterized by θ^Q and θ^μ, respectively:

$$\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\theta^{Q'}$$

$$\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\theta^{\mu'}$$

- A replay buffer *R* is used to conduct the update
- Minimize the loss, L, over the value-function network parameters

$$L(\theta^Q) = \frac{1}{N}\sum_i \left(y_i - Q(s_i, a_i|\theta^Q)\right)^2$$

$$y_i = r_i + \gamma Q'\left(s_{i+1}, \mu(s_{i+1}|\theta^{\mu'})|\theta^{Q'}\right)$$

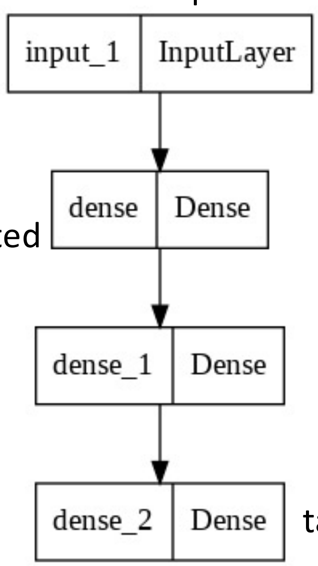
- The exploration noise for DDPG is commonly considered as an Orstein-Uhlenbeck process

$$x_{k+1} = x_k + \eta(\mu - x_k)\Delta t + \sigma\sqrt{\Delta t}\mathcal{N}(\vartheta, \sigma^2)$$

Network Architectures

- Algorithms are implemented in Python using ML packages from TensorFlow / Keras

State Input



Fully Connected ReLU

Actor Network

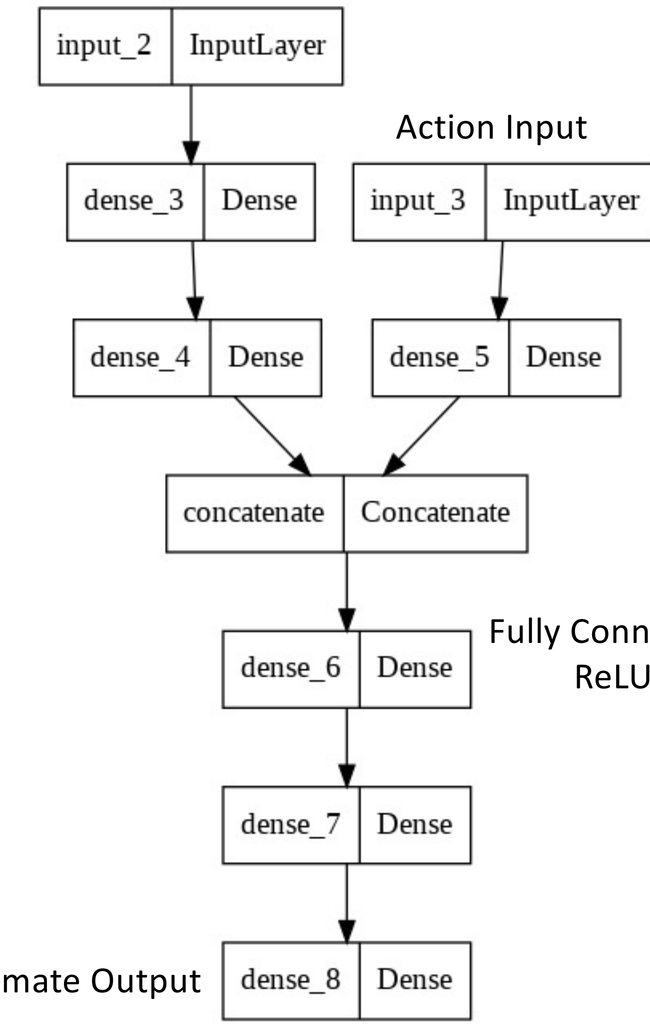
tanh or sigmoid

Action Output

Critic Network

State Input

Action Input



Fully Connected ReLU

Q Estimate Output

West Virginia University.
BENJAMIN M. STATLER COLLEGE OF ENGINEERING AND MINERAL RESOURCES

Algorithm

Algorithm 7: DDPG for Offset-Free Control

Input: Initialize critic and actor networks as $Q^U\left(s, a \middle| \theta^{Q^U}\right)$, $Q^K\left(s, a \middle| \theta^{Q^K}\right)$, $\mu^U\left(s \middle| \theta^{\mu^U}\right)$, $\mu^K\left(s \middle| \theta^{\mu^K}\right)$

Initialize noise object $\mathcal{N}$

Set the target network weights $(\theta^{Q^{U\prime}}, \theta^{\mu^{U\prime}}) \leftarrow (\theta^{Q^U}, \theta^{\mu^U})$, $(\theta^{Q^{K\prime}}, \theta^{\mu^{K\prime}}) \leftarrow (\theta^{Q^K}, \theta^{\mu^K})$

Initialize the replay buffer R

Set ε

for ep = 1:Max$_{ep}$

    Set $s = s^0 \in S$

    Initialize noise objects $\mathcal{N}_1, \mathcal{N}_2$

    for k = 1:Max$_k$

        Get $a_k = \pi(s_k) = \begin{cases} \mu\left(s_k \middle| \theta^{\mu^U}\right) + \mathcal{N}_1, & if\ \|x\| > \varepsilon_k \\ \mu\left(s_k \middle| \theta^{\mu^K}\right) + \mathcal{N}_2, & else \end{cases}$

Update Actor Network

Update Target Networks

Update Critic Network

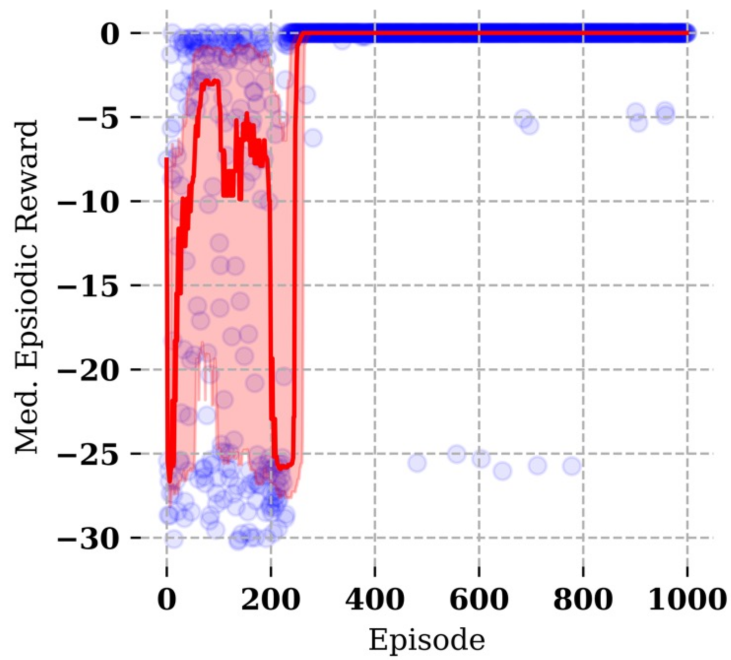
Noise for exploration

Target networks for learning stability

Apply action to plant

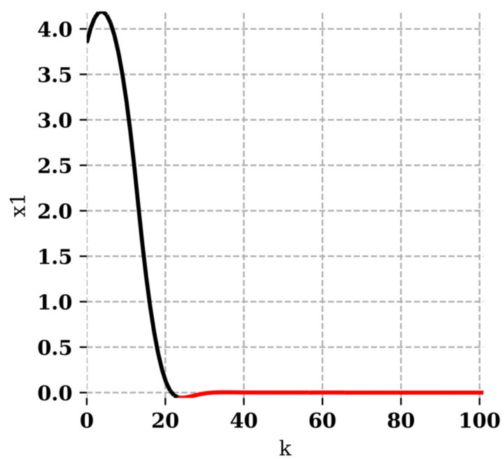
Replay buffer for learning over multiple samples

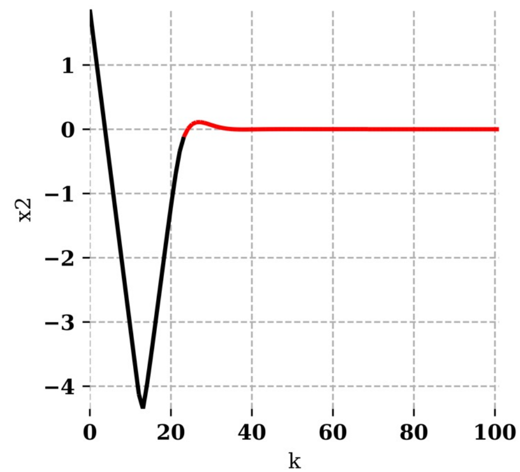
Results – Linear System

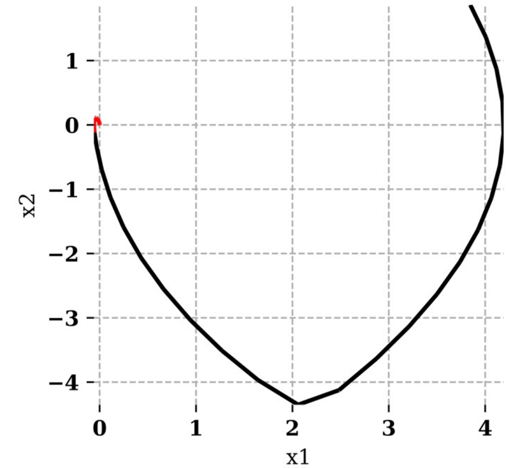


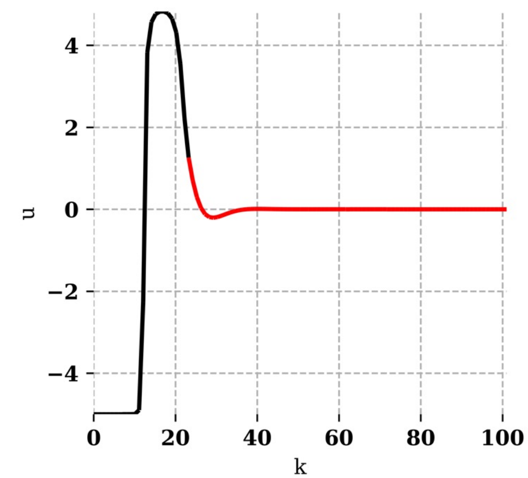
$$A = \begin{bmatrix} 1 & 0.1 \\ 0 & 1 \end{bmatrix},\ B = \begin{bmatrix} 0.005 \\ 0.1 \end{bmatrix},\ C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

Results – Linear System



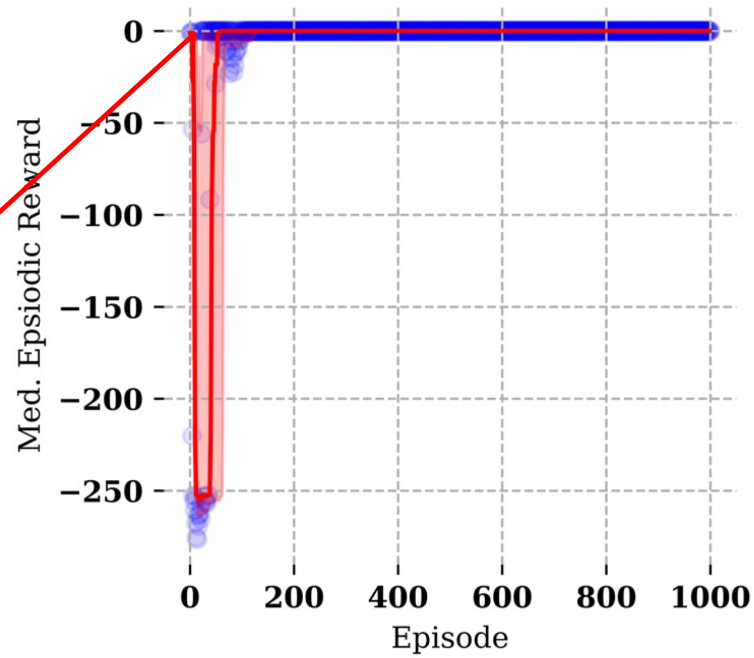




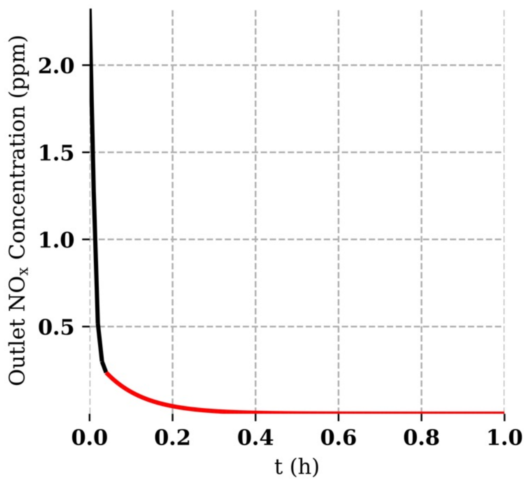


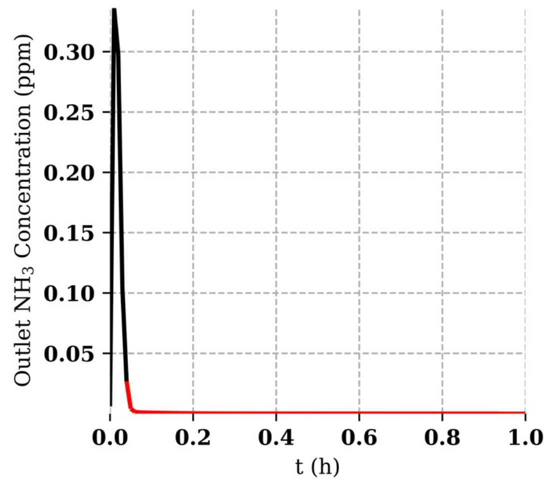
Results – SCR Control

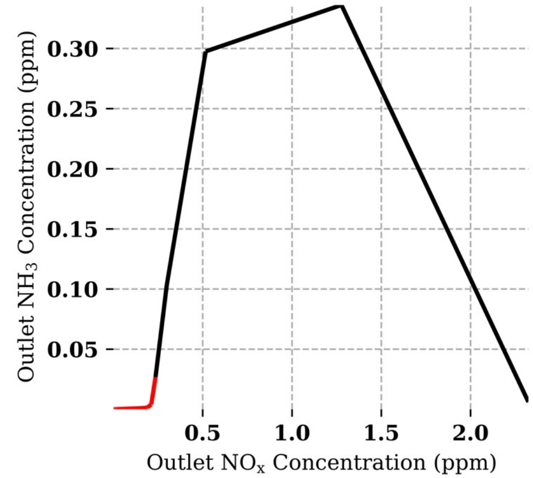
Initial reward higher than
double integrator

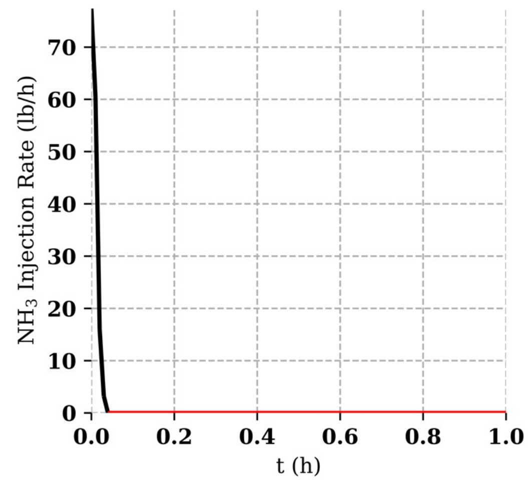


Results – SCR Control









WestVirginiaUniversity.
BENJAMIN M. STATLER COLLEGE OF
ENGINEERING AND MINERAL RESOURCES

Conclusions

- RL-MPC algorithm developed for selecting control parameters online
- Results presented for an industrially relevant example
  - Superior performance achieved with offline learning, further improved with learning on the online system
- Developed algorithms for offset-free actor-critic control and applied to linear and nonlinear systems
- There exist considerable opportunities to exploit strengths of RL for improving the performance of control systems by using RL by itself under sufficient performance guarantee or in combination with existing controllers

Acknowledgements

- My student Dr. Elijah Hedrick (now in GE) conducted most of the research presented in this work. My student Katherine Hedrick contributed to model development.
- Thanks to Dr. Stephen Zitney, and Dr. Benjamin Omell, NETL for their valuable contributions
- The authors would like to acknowledge funding from the U.S. Department of Energy's National Energy Technology Laboratory under the Mission Execution and Strategic Analysis contract (DE-FE0025912) for support services through KeyLogic Systems, Inc. under P.O. 5000-074.

West Virginia University.
BENJAMIN M. STATLER COLLEGE OF
ENGINEERING AND MINERAL RESOURCES

Thank you!

Questions?