The 2023 IEEE 12th Data Driven Control and Learning Systems Conference

Q-learning based optimal tracking control of coal-fired power plants

Presenter: mengjun Yu

Author: xiaomin Liu, mengjun Yu, chunyu yang, linna zhou

Email: xiaominliu@cumt.edu.cn, mengjunyu@cumt.edu.cn

2023-5-14

CONTENTS

1 Research Background and Significance

2 Problem Description

3 Algorithm Design

4 Numerical Simulations

5 Conclusion

1

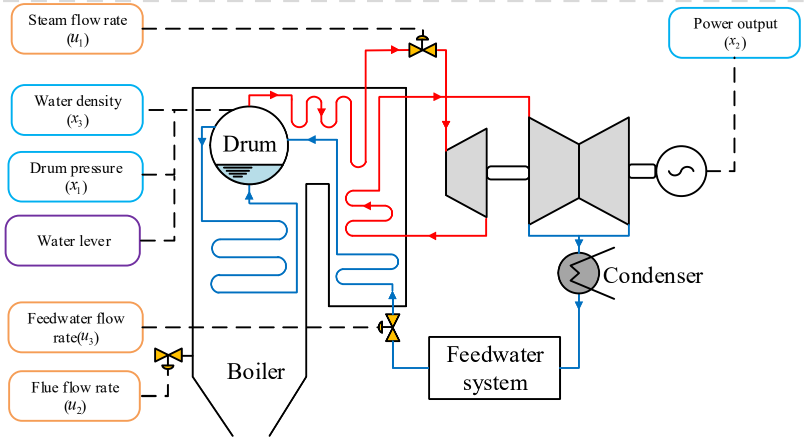
Research Background and Significance

1 Research background and significance

Optimal control of load tracking in coal-fired power generation

- **Load tracking** the ability of a generator set to adjust its output in response to changes in grid load

- **Optimal control** is a technique that uses mathematical optimization methods to design controllers, can define an objective function to quantitatively describe the performance of the controller, and then solve the controller through optimization methods. The background of the optimal control problem of load tracking in coal-fired power generation is to achieve efficient, energy-saving and flexible operation of coal-fired generator sets, while meeting the load demand and safety and stability requirements of the power grid



1 Research background and significance

Shortcomings of traditional model-based and data-based

Challenges of optimizing the operation of coal-fired power generation systems

- An accurate model
- Poor adaptive ability
- The amount of data required is large
- Asymmetric control constraints

An adaptive dynamic programming learning method based on Q-learning is proposed

- A mode-free, off-policy adaptive artificial intelligence algorithm
- Without adding additional objective function penalty terms, the asymmetric control constraint is converted to solve the input constraint asymmetry problem
- Improve data utilization with experience playback

2

Problem Description

2 Problem description

□ **Consider a general nonaffine nonlinear discrete-time system:**

$$x(k+1)=f\big(x(k),u(k)\big)\quad(1)$$

□ **Desired reference trajectory:**

$$r(k+1)=h\big(r(k)\big)\qquad(2)$$

□ **Denoting the tracking error as $\boldsymbol{e}(\boldsymbol{k})\triangleq\boldsymbol{x}(\boldsymbol{k})-\boldsymbol{r}(\boldsymbol{k})$, the error system is represented：**

$$e(k+1)=f\big(e(k)+r(k),u(k)\big)-h\big(r(k)\big)(3)$$

Remark:

- f,h,F Unknown,$f(0,0)=0$
- r is a bounded reference signal， $h(r)$ is a Lipschitz continuous vector function
- $0<\gamma\leqslant 1$is the discount factor, $W(e)$ and $R(u)$ are positive definite functions

□ **The state of the augmentation system is defined as:$\boldsymbol{y}(\boldsymbol{k})\triangleq[\boldsymbol{e}^{\top}(\boldsymbol{k})\quad \boldsymbol{r}^{\top}(\boldsymbol{k})]^{\top}$, then the augmentation system:**

$$y(k+1)=F\big(y(k),u(k)\big)\quad(4)$$

□ **Considering the optimal tracking control problem for a given target value, design the discount performance metrics as follows:**

$$J(y(0),u)\triangleq\sum_{l=0}^{\infty}\gamma^{l}\mathcal{R}\big(y(l),u(l)\big)=\sum_{l=0}^{\infty}\gamma^{l}[W(e)+R(u)]\quad(5)$$

3 Algorithm Design

3.1 Asymmetric input constraint system

□ Asymmetric inputs are symmetrically transformed around the median of the control range

Suppose the actual control input v of the system is constrained to

$$\bar{v}_{\min,j} \leq v_j \leq \bar{v}_{\max,j}\ , j = 1,2,\ldots,m \qquad (6)$$

The length of the constraint interval that controls the input is defined as

$$\bar{v}_{z,j} = \frac{(\bar{v}_{\max,j} - \bar{v}_{\min,j})}{2}, j = 1,2,\ldots,m \qquad (7)$$

Let $\bar{V}_z \in \mathbb{R}^{m\times m}$ be a constant diagonal matrix given by the following equation

$$\bar{V}_z = \begin{bmatrix} \bar{v}_{z,1} & 0 & \cdots & 0 \\ 0 & \bar{v}_{z,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \bar{v}_{z,m} \end{bmatrix} \qquad (8)$$

The median value of the controllable interval as

$$\bar{v}_{d,j} = \frac{(\bar{v}_{\max,j} + \bar{v}_{\min,j})}{2}, j = 1,2,\ldots,m \qquad (9)$$

The unconstrained asymmetric input u performs symmetry constraints

$$\bar{v} = \bar{V}_z \varphi\left(\bar{V}_z^{\ -1}(u - \bar{v}_{d,j})\right) \qquad (10)$$

Note: The asymmetric input is symmetrically converted because the general constraint is based on symmetric constraint processing

$\varphi(\cdot)$ take $tanh(\cdot)$, $|\varphi(\cdot)| \leq 1$, here to get $\bar{v}_{\max,j} \leq v = \bar{v} + \bar{v}_{d,j} \leq \bar{v}_{\min,j}$

3.1 Asymmetric input constraint system

□ Asymmetric control constraint problem transformation

For a constraint control system, the system can be represented as

$$x(k+1) = f(x(k), v) \qquad (11)$$

The constraint control strategy corresponds to the transformation of the augmentation system

$$y_v(k+1) = \begin{bmatrix} f(e(k)+r(k), v) - h(r(k)) \\ h(r(k)) \end{bmatrix} \qquad (12)$$

The cost function designed without additional penalties is as follows

Note: No additional penalties are reflected in

$$R(v) = R\left(\bar{v}_{d,j} + \bar{V}_z \varphi\left({\bar{V}_z}^{-1}(u - \bar{v}_{d,j})\right)\right),$$

The general treatment is

$$R(v) \\ = R\left(\bar{v}_{d,j} + 2\bar{V} \sum_{i=1}^{m} \int_{\bar{u}_{d,j}}^{v_i} \varphi^{-1}\left({\bar{V}_z}^{-1}(v - \bar{v}_{d,j})\right) dv\right)$$

$$J(y(0), u) \triangleq \sum_{l=0}^{\infty} \gamma^l \mathcal{R}(y(l), u(l)) = \sum_{l=0}^{\infty} \gamma^l [W(e) + R(u)] \qquad (13)$$

The unconstrained optimal control corresponds to： $u^*(y) \triangleq arg\,\min_u J(y_v(0), u)$

The constraint optimal control design is: $v^*(y) = \bar{v}_{d,j} + \bar{V}_z \varphi\left({\bar{V}_z}^{-1}(u^* - \bar{v}_{d,j})\right)$

The performance indicators represented by unconstrained and constrained are minimized at the same time, so the constrained optimal control problem is transformed into an unconstrained optimal control problem

3.2 Critic-only Q-learning control algorithm

□ Optimal control problem solving

Bellman's equation for introducing the state value function:

$$\begin{aligned} V_u(y(k)) &= \mathcal{R}(y(k),u(k)) + \sum_{l=k+1}^{\infty} \gamma^{l-k}\mathcal{R}(y_v(l),v(l)) \\ &= \mathcal{R}(y(k),u(k)) + \gamma \sum_{l=k+1}^{\infty} \gamma^{l-(k+1)}\mathcal{R}(y_v(l),v(l)) \\ &= \mathcal{R}(y(k),u(k)) + \gamma V_u(y_v(k+1)). \end{aligned} \tag{14}$$

The Bellman equation for the optimal state value function satisfies

$$V^*(y(k)) = \min_u\{\mathcal{R}(y(k),u(k)) + \gamma V^*(y(k+1))\} \tag{15}$$

The optimal control strategy is

$$u^*(y) = \arg\min_u V^*(y) \tag{16}$$

Remark:

- $V_u(0) = 0$
- $V^*(y) \triangleq V_{u^*}(\mathrm{y}) = \min_u V_u(\mathrm{y})$

3.2 Critic-only Q-learning control algorithm

□ Optimal control problem solving

Introduce state-action functions

$$Q_u(y(k),a) = \mathcal{R}(y(k),a) + \gamma Q_u\big(y(k+1),u(k+1)\big) \tag{17}$$

Iterative algorithm based on Q-learning strategy

(Policy evaluation)

$$Q^{(i)}(y(k),a) = \mathcal{R}(y(k),a) + \gamma Q^{(i)}\left(y(k+1),u^{(i)}(k+1)\right) \tag{18}$$

(Policy improvement)

$$u^{(i+1)}(y) = u^{(i)}(y) - \alpha \frac{\partial Q^{(i)}(y,a)}{\partial a}\bigg|_{a=u^{(i)}(y)} \tag{19}$$

Algorithm 1: Q-learning

```
1: Let $u^{(0)}(y)$ be an initial admissible control policy, and i = 0;
2: (Policy evaluation) Solve the equation
      $Q^{(i)}(y(k),a) = \mathcal{R}(y(k),\mathrm{a}) + \gamma Q^{(i)}\left(y(k+1),u^{(i)}(k+1)\right)$
3: (Policy improvement) Solve the equation
      $u^{(i+1)}(y) = u^{(i)}(y) - \alpha \frac{\partial Q^{(i)}(y,a)}{\partial a}\Big|_{a=u^{(i)}(y)}$
4: Let $i = i + 1$, If $u$ converges, stop the iteration, else go back to step 2 and continue
```

3.2 Critic-only Q-learning control algorithm

□ Neural network implementation

Use neural networks Q function $Q(y, a)$

$$\hat{Q}(y, a) = \sum_{j=1}^{L} \theta_j \psi_j(y, a) = \Psi_L^{\top}(y, a)\hat{\theta} \qquad (20)$$

The residual error due to the Cirtic neural networks approximation errors

$$\begin{aligned}\epsilon^{(i)}(y, a) &= \hat{Q}^{(i)}(y, a) - \gamma\hat{Q}^{(i)}\left(y', \hat{u}^{(i)}\right) - \mathcal{R}(y, a) \\ &= \left[\Psi_L(y, a) - \gamma\Psi_L\left(y', \hat{u}^{(i)}\right)\right]^{\top}\hat{\theta}^{(i)} - \mathcal{R}(y, a)\end{aligned} \qquad (21)$$

The residuals as follows are minimized

$$min \sum_{l=1}^{M} \left(\epsilon_{[l]}^{(i)}\right)^2 \qquad (22)$$

The update $\hat{\theta}$ by least squares is as follows

$$\hat{\theta}^{(i)} = \left[\left(Z^{(i)}\right)^{\top} Z^{(i)}\right]^{-1} \left[Z^{(i)}\right]^{\top} \eta \qquad (23)$$

3.3 The Asymmetric Constraints Q-learning with Experience Replay

Algorithm 1: The Asymmetric Constraints Q-learning with Experience Replay

```
1: Initialize the coal-fired generation data buffer $D$, set the capacity size to$N$
2: Random initial network parameter $\theta$, initialize the state action value function $Q$, and set the coal-fired power generation status feature vector $\Psi$
3: for epoch 1 to $M$
4:   Initialize the random noise $\mathcal{N}$of the coal-fired power generation system and a feasible initial control input $u_0$
5:   for $t = 1$ to T
6:     Observe the initial state of coal-fired power generation $y_t$
7:     Based on gradient descent, select the unconstrained control input action
           $u_t = u_{t-1} - \alpha \frac{\partial \hat{Q}(y,a)}{\partial a}\Big|_{a=u(y)} + \mathcal{N}_t$
8:     Ensure that the coal-fired power generation system performs control input $a_t$within the constraint range
       according to the constraint formula, calculate the reward $\mathcal{R}_t$ and observe the next sampling point status of
       the coal-fired power generation system $y_t'$
9:   (Data store) Save the real-time collection of coal-fired power generation data tuples $\{y_{[t]}, a_{[t]}, y'_{[t]}, \mathcal{R}_{[t]}\}$in
       buffer $D$, and discard the oldest data if $D$ exceeds the capacity size $N$.
     End for
10:  Randomly sampled minimum batch data transition $\{y_{[k]}, a_{[k]}, y'_{[k]}, \mathcal{R}_{[k]}\}$from coal-fired power generation
           historical data $D$
11:  (Policy evaluation) Minimize the residual error $\epsilon$ by least squaresand update the network parameter $\hat{\theta}$
           $\hat{\theta}^{(i)} = \left[\left(Z^{(i)}\right)^{\top} Z^{(i)}\right]^{-1} \left[Z^{(i)}\right]^{\top} \eta$
12:  (Policy improvement) Based on one-step gradient descent:
           $\hat{u}^{(i+1)}(y) = \hat{u}^{(i)}(y) - \alpha \frac{\partial \hat{Q}^{(i)}(y,a)}{\partial a}\Big|_{a=\hat{u}^{(i)}(y)}$
           $= \hat{u}^{(i)}(y) - \alpha \frac{\partial \Psi_L^{\top}(y,a)}{\partial a}\Big|_{a=\hat{u}^{(i)}(y)} \hat{\theta}^{(i)}$
13:  For very small arguments $\varepsilon > 0$, stop iteration if $\|\hat{\theta}^{(i)} - \hat{\theta}^{(i-1)}\| \leqslant \varepsilon$; Otherwise, $i = i + 1$, go back to
         step 11 to continue.
14: end for
```

4

Numerical Simulations

4 Coal-fired power generation system

Discrete systems for coal-fired power generation

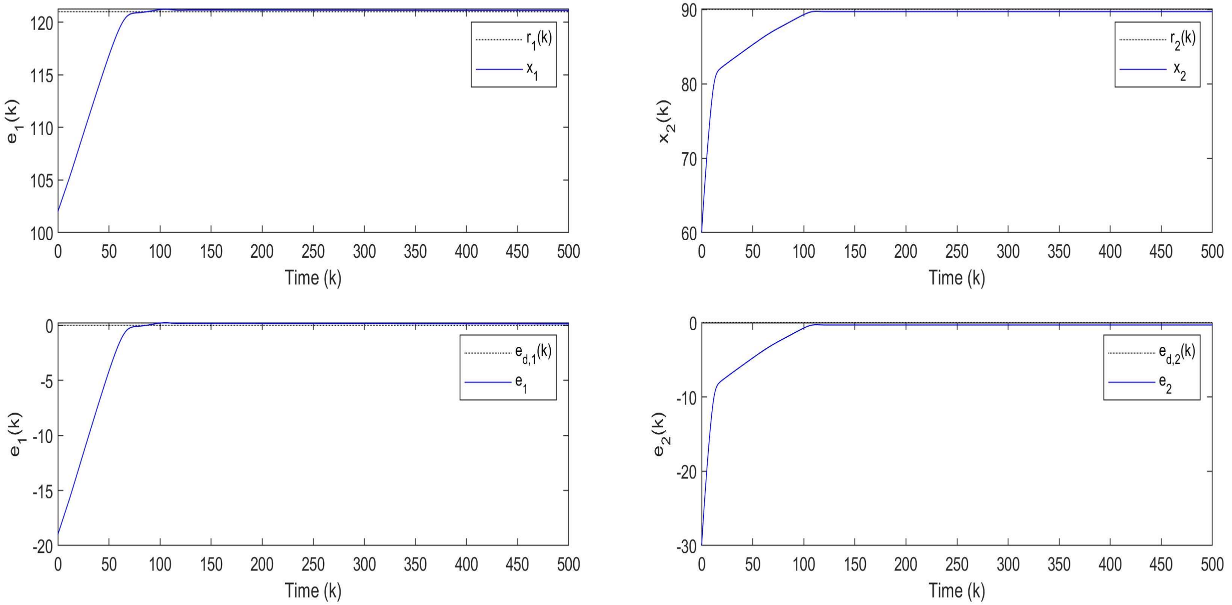
$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix} = \begin{bmatrix} x_1(k) \\ -0.016T_s x_1(k)^{\frac{9}{8}} + (1-0.1T_s)x_2(k) \\ 0.0022T_s x_1(k) + x_3(k) \end{bmatrix}$$

$$+ \begin{bmatrix} 0.9T_s & -0.0018T_s x_1(k)^{\frac{9}{8}} & -0.15T_s \\ 0 & 0.073T_s x_1(k)^{\frac{9}{8}} & 0 \\ 0 & -0.0129T_s x_1(k) & 1.6588T_s \end{bmatrix}$$

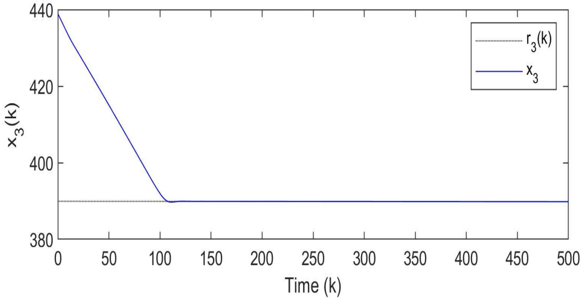
$$\times \begin{bmatrix} u_1(k) \\ u_2(k) \\ u_3(k) \end{bmatrix}$$

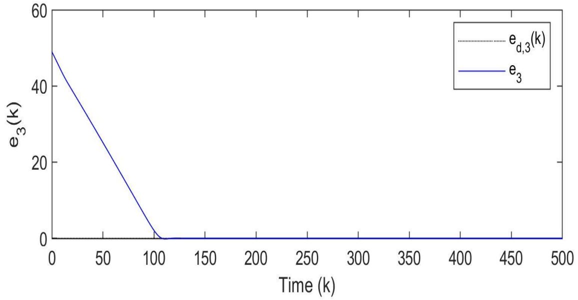
Sampling interval Ts = 0.5 seconds. The initial and target states are $x(0) = [102,60,438.93]^{\top}$ and $x_t = [121,90,389.92]^{\top}$ respectively. The initial and steady-state control inputs $u(0) = [0.3102,0.6711,0.3967]^{\top}$ and $u_d = [0.4385,0.7787,0.5720]^{\top}$, respectively. Then, from the control constraint, it can be concluded that the bounded $u_z(k)$ of the control input is $\bar{u}_z = [1.0,1.0,1.0]^{\top}$, $\gamma = 0.99$

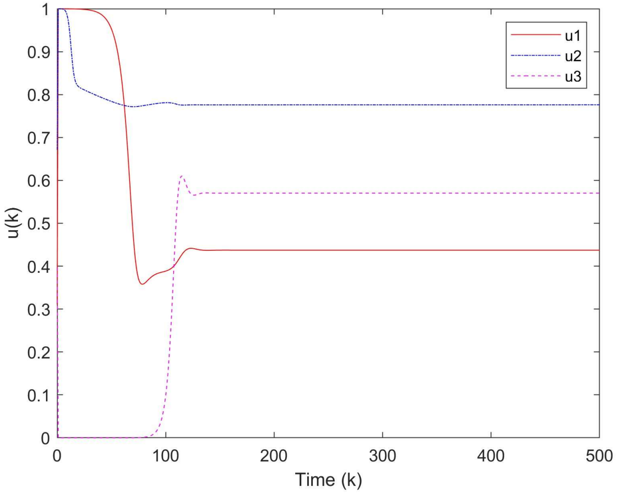
4 Coal-fired power generation system



4 Coal-fired power generation system







5 Conclusion

5 Conclusion

In this paper, a Q-learning algorithm based on critic-only structure is proposed to solve the optimal load tracking control problem of model-free, non-affine, nonlinear discrete coal-fired power generation system with asymmetric input constraints, and an experience replay technique is introduced to improve sample utilization. The simulation results show that the proposed algorithm has good stability and tracking effect.

In future work, consider further improving the performance and efficiency of algorithms by introducing deep reinforcement learning and other advanced technologies. We will also focus on the possibility of integrating the proposed algorithm with a real-time data acquisition and control system to enable online adaptive control of coal-fired power plants.

Thank you