A Bias Compensating Value Iteration Based Q-learning Algorithm for Model-Free Game-Theoretic HVAC Optimal Control

Junaid Anwar[1] and Syed Ali Asad Rizvi[2]

[1]EE, San José State University, USA

[2]EE, Tennessee Technological University, USA

Paper ID: MoAT8.1
Session: Poster Session, Grand Station I-II
Presenter: Junaid Anwar

October 6, 2025

HVAC Game Control Formulation

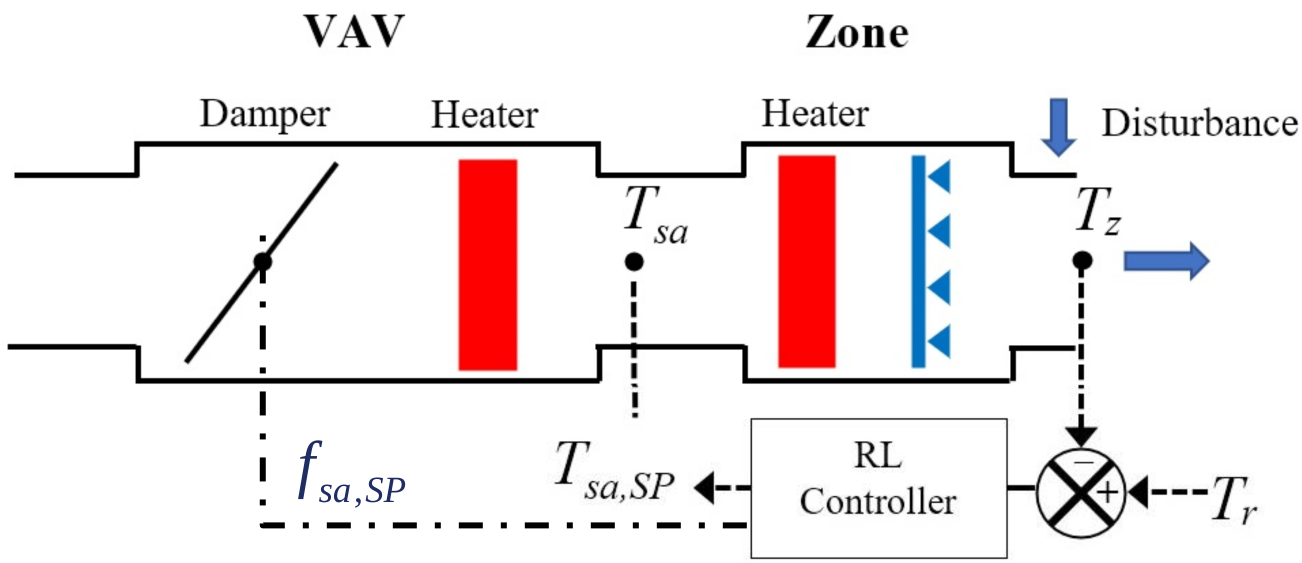


Fig. 1 An RL loop for a zone controlled with two players as decision variables

- Unlike classical LQR, here each player has its own distinct cost function

$$J_i = \sum_{k=0}^{\infty} \left(e^\top(k)\, Q_{i,e}\, e(k) + w^\top(k)\, Q_{i,w}\, w(k) + \sum_{j=1}^{2} u_j^\top(k)\, R_{ij}\, u_j(k) \right)$$

Individual player weights

State Feedback Two-Player Game Q-functions

- Game theoretic control policy for each player

$$u_i(k) = -K_i \begin{bmatrix} x^\top(k) & w(k)^\top \end{bmatrix}^\top$$

Optimal game policy

- Individual Q-function

$$Q_i(z) \triangleq z^\top(k)\, H_i\, z(k), \quad i = 1, 2 \text{ player}$$

$$z(k) = \begin{bmatrix} X^\top(k) & u_1(k) & u_2(k) & c \end{bmatrix}^\top$$

Bias compensation term

- Our recent state feedback PI method needs an initially stabilizing policy

IJRNC ’25

NZS Q-learning VI Algorithm

Algorithm: A Two-Player Non-Zero-Sum Game Q-learning Value Iteration (VI) Algorithm with Bias Compensation

input: input-state data
output: H_i^*
1. **initialize.** Select any initial policies $u_1^0(k)$ and $u_2^0(k)$ with exploration signals. Set $j \leftarrow 0$.
2. **acquire data.** Apply input $u_1^0(k)$ and $u_2^0(k)$ to collect $L \geq l(l+1)/2$ datasets of $\{x(k), w(k), u_1(k), u_2(k)\}$.
3. **repeat**
4. **Value update.** For each player $i = 1, 2$, learn the solution of the data-driven coupled game Bellman equations:

$$z^\top(k)H_i^{j+1}z(k) = x^\top(k)Q_{i,x}x(k)+w^\top(k)Q_{i,w}w(k)+\sum_{m=1}^{2} u_m^\top(k)R_{ij}u_m(k)+z^\top(k+1)H_i^j z(k+1)$$

5. **policy improvement.** For each player $i = 1, 2$, determine an improved policy from the coupled gain equations:

$$K_1^{j+1} = \left(I - (H_{1,u_1u_1}^{j+1})^{-1}H_{1,u_1u_2}^{j+1}(H_{2,u_2u_2}^{j+1})^{-1}H_{2,u_2u_1}^{j+1}\right)^{-1}(H_{1,u_1u_1}^{j+1})^{-1}\times\left(H_{1,u_1X}^{j+1} - H_{1,u_1u_2}^{j+1}(H_{2,u_2u_2}^{j+1})^{-1}H_{2,u_2X}^{j+1}\right)$$

$$K_2^{j+1} = \left(I - (H_{2,u_2u_2}^{j+1})^{-1}H_{2,u_2u_1}^{j+1}(H_{1,u_1u_1}^{j+1})^{-1}H_{1,u_1u_2}^{j+1}\right)^{-1}(H_{2,u_2u_2}^{j+1})^{-1}\times\left(H_{2,u_2X}^{j+1} - H_{2,u_2u_1}^{j+1}(H_{1,u_1u_1}^{j+1})^{-1}H_{1,u_1X}^{j+1}\right)$$

6. $j \leftarrow j + 1$

until $\left\|K_i^j - K_i^{j-1}\right\| < \varepsilon$ for some small $\varepsilon > 0$.

Tracking Performance

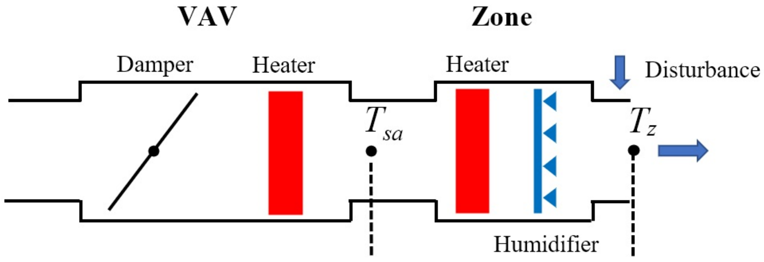


Fig. 1: An HVAC zone with two players as decision variables

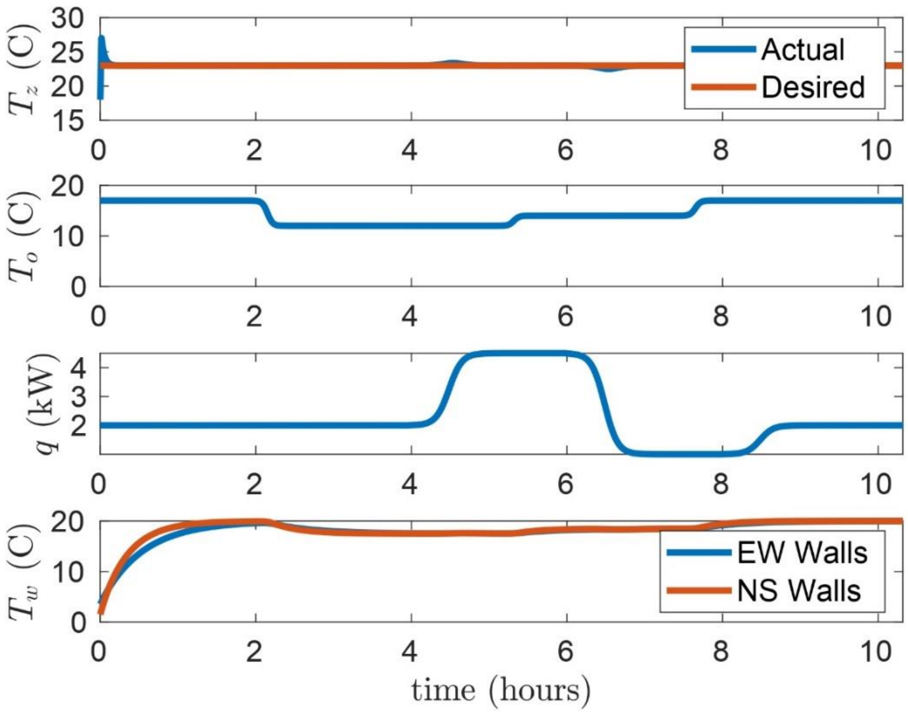
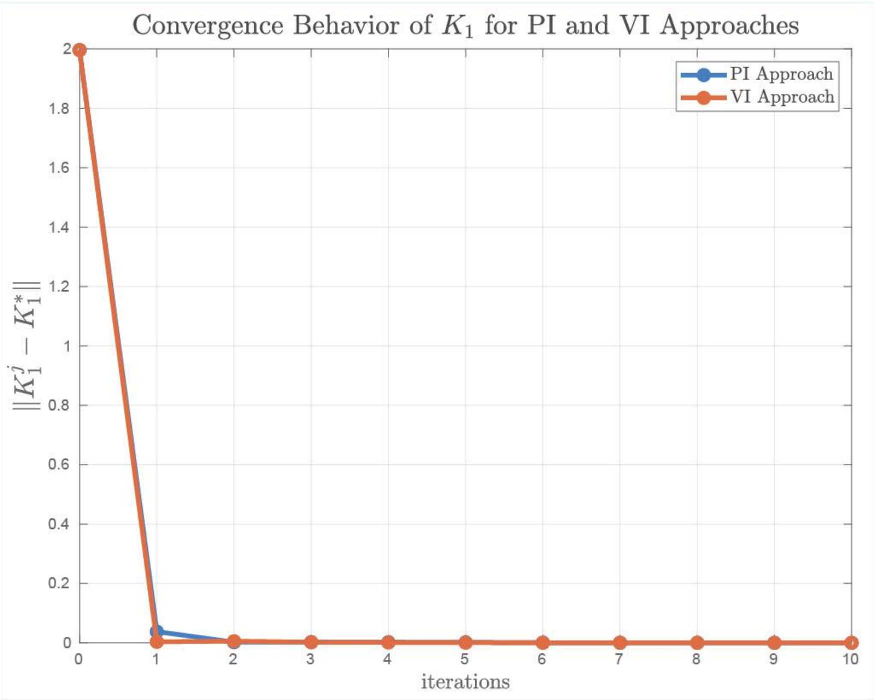


Fig. 2: Tracking response of the learned controller

Parameter Convergence



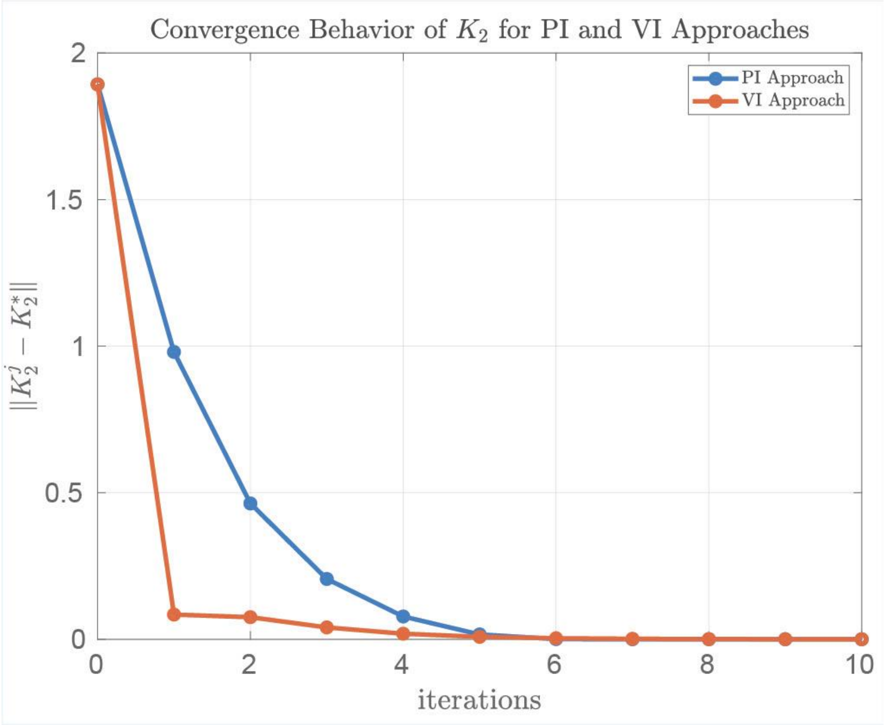


Fig. 3 Convergence of Player 1 and 2 gains under varying load and disturbances

Benefits of proposed VI method

No warm start: learns without a stable policy; simpler updates

GARE-optimal: reaches model-based state-feedback solution

Join me in the interactive session to learn more about this work.

Thank you!