African Journal of Advanced Pure and Applied Sciences (AJAPAS)
Online ISSN: 2957-644X
Volume 4, Issue 1, January - March 2025
Page No: 513-519
Website: https://aaasjournals.com/index.php/ajapas/index
ISI 2024: 0.877 **SJIFactor 2024: 6.752** **معامل التأثير العربي: 1.62**

Weighted Policy Learning Based Control for Two-Tank Level System

Mostafa D. Awheda [1 *], Saad A. Abobakr [2]
[1,2] Department of Control Engineering, College of Electronic Technology, Bani Walid, Libya

التحكم القائم على التعلم المستند إلى السياسة المرجحة لنظام مستوى الخزانين

مصطفى ضي أوحيدة [*1]، سعد عبد الباسط أبوبكر [2]
[2،1] قسم هندسة التحكم الآلي، كلية التقنية الإلكترونية، بني وليد، ليبيا

[*]Corresponding author: mdawheda@gmail.com

Received: February 12, 2025	Accepted: March 23, 2025	Published: March 27, 2025

Abstract:
Reinforcement learning (RL) is a model-free framework in which agents learn control policies through trial-and-error interaction with the environment. While classical controllers such as PID, LQR, and MPC guarantee stability and interpretability, they tend to rely on accurate models and become suboptimal in nonlinear and uncertain system dynamics situations. We address the two-tank fluid level control benchmark, with its high coupling and nonlinear outflows, in this work using the multi-agent RL formulation by the Weighted Policy Learning (WPL) algorithm. The level of the second tank is controlled by WPL's policy-gradient weighting to ensure smooth convergence under non-stationarity. Simulation results demonstrate rapid setpoint tracking, minimal overshoot, and higher disturbance robustness and reflect the effectiveness of, as well as innovativeness in, applying WPL to process control issues.

Keywords: Reinforcement Learning, Weighted Policy Learning, Two-Tank Level Control.

الملخص
التعلم التعزيزي (RL) هو إطار عمل خالٍ من النماذج، حيث يتعلم الوكلاء سياسات التحكم من خلال التفاعل مع البيئة بطريقة التجربة والخطأ. في حين تضمن وحدات التحكم الكلاسيكية، مثل PID وLQR وMPC، الاستقرار وقابلية التفسير، إلا أنها تميل إلى الاعتماد على نماذج دقيقة وتصبح دون المستوى الأمثل في حالات ديناميكيات النظام غير الخطية وغير المؤكدة. نتناول في هذا العمل معيار التحكم في مستوى السوائل بخزانين، بما يتميز به من اقتران عالٍ وتدفقات خارجية غير خطية، باستخدام صياغة التعلم التعزيزي متعدد الوكلاء بواسطة خوارزمية تعلم السياسات الموزونة (WPL). يتم التحكم في مستوى الخزان الثاني بواسطة ترجيح تدرج السياسات في WPL لضمان تقارب سلس في ظل عدم الثبات. تُظهر نتائج المحاكاة تتبعًا سريعًا لنقطة الضبط، وتجاوزًا ضئيلًا، ومقاومة أعلى للاضطراب، وتعكس فعالية، فضلًا عن الابتكار، في تطبيق WPL على مشكلات التحكم في العمليات.

الكلمات المفتاحية: التعلم التعزيزي، التعلم بالسياسة المرجحة، التحكم على مستوى الخزانين.

Introduction

Reinforcement learning (RL) is a learning framework in which an agent learns to transform situations into actions through interacting with the environment and enhancing its performance from the resulting feedback [1]. RL has been widely applied and has gained much attention in intelligent robotic control systems [5]–[8]. It has also been demonstrated to be an effective tool for solving nonlinear optimal control problems [9]. Reinforcement learning (RL) is a paradigm for learning whereby an agent improves its action by interacting with its environment. At every

discrete time instant, the agent senses the present state, chooses an action, and causes the environment to evolve into a new state. A scalar reward is then received, quantifying how valuable the new transition is. The goal of the agent is to gain a policy that maximizes the long-run expected cumulative reward [1], [10].

Reinforcement learning (RL) is a powerful model-free control approach that enables autonomous tuning of parameters through trial-and-error interactions driven by the controller's exploration of the environment [1]. The combination of reinforcement learning algorithms with classical control techniques is a potential solution for the development of stable and efficient control systems. Classical controllers such as PID, LQR, or MPC offer stability and interpretability and RL offers adaptability in dynamic or uncertain environments where classical methods may fall short. Dev et al. [11] integrated classical control theory and reinforcement learning approaches, emphasizing the generality and utility of machine learning algorithms in matters concerning control. Puriel et al. [12] proposed a method in which the robot's PID controller is enhanced through compensation using reinforcement learning techniques. Bałazy et al. [13] developed a novel reinforcement learning (RL) algorithm that employs a policy designed to be robust against object nonlinearity. Lee et al. [14] proposed a near-optimal semi-active suspension ride comfort controller based on deep reinforcement learning. Yaghmaie et al. [15] proposed a model-free reinforcement learning approach to utilize linear quadratic control and demonstrated the potential of RL in optimal control problems without requiring an explicit system model.

Two tank fluid level control is a classic process control benchmark problem with nonlinear dynamics and significant coupling effects [16], [17]. Standard PID or model-based controllers are normally developed on the basis of precise system modeling and may underperform in the presence of disturbances or parameter variations [2]. Reinforcement learning (RL), particularly in the multi-agent case, is an appealing alternative since it has the ability to learn directly from experience in the environment the control policies [1], [5]. In this paper, we propose a new application of the Weighted Policy Learning (WPL) algorithm—a multi-agent RL method—to the two-tank level control problem. WPL dynamically adjusts the learning rate of agents in the context of policy gradients, ensuring smooth and stable convergence for non-stationary multi-agent systems. The approach formulates each tank as an independent learning agent, enabling decentralized and adaptive control, thereby dealing with tanks' inherent coupling. The formulation improves not only the robustness and effectiveness of control but also offers a new framework for the generalization of game-theoretic MARL approaches to process control problems.

The Weighted Policy Learning (WPL) Algorithm

The Weighted Policy Learning (WPL) algorithm [4] employs a policy-gradient approach to update each agent's strategy. In WPL, all players are assumed to have access to the value function of the game, which is used to compute the policy gradient $\delta(s_t, a)$. WPL has been proven to converge to Nash equilibria in standard two-player, two-action games and in several larger game structures. Remarkably, WPL requires minimal information: each agent needs only its own received reward for the chosen action, without knowledge of other agents' actions, rewards, or the game's payoff matrices. Furthermore, WPL does not require knowing the Nash equilibrium in advance. During learning, WPL adjusts its update speed based on the sign of the policy gradient. It adapts quickly when the gradient reverses direction and more conservatively when the gradient maintains its sign. The sign of the policy gradient is determined via:

$$\hat{\delta}(s_t, a) \;=\; Q(s_t, a) - V(s_t),$$

where $Q(s_t, a)$ represents the expected value of taking the action a in state s_t and $V(s_t)$ represents the state's average reward.

The learning agent uses the following equation to update its Q-table:

$$Q_{t+1}(s, a_t) = (1 - \theta)Q_t(s, a_t) + \theta\left[r_t + \zeta \max_{a'} Q_t(s', a')\right] \qquad (1)$$

Here, t denotes the number of times the state s has been visited, θ represents the learning rate, r_t is the immediate reward received by the agent at state s, a_t is the action selected by the agent in state s, and ζ denotes the discount factor.

The learning agent updates its Q-table and policy over time. The Q-table is updated as in the previous equation, Eq. (1), and the policy is updated using the following rule [4]:

$$\pi_{t+1}(s_t, a) = \pi_t(s_t, a) + \eta\delta(s_t, a) \qquad (2)$$

and,

$$\pi_t(s_{t+1}) = \text{limit}\big(\pi_t(s_{t+1})\big)$$

Here's how the policy gradient $\delta(s_t, a)$ is calculated [4]:

$$\delta(s_t, a) = \begin{cases} \hat{\delta}(s_t, a).\big(1 - \pi_t(s_t, a)\big) & if\ \hat{\delta}(s_t, a) > 0 \\ \hat{\delta}(s_t, a).\pi_t(s_t, a) & otherwise \end{cases} \qquad (3)$$

where:

η is a learning rate such that $\eta \in (0,1)$, $V(s_t) = \sum_{a \in A} \pi_t(s_t, a) Q_t(s_t, a)$ is the average reward at state s_t, and $limit(\pi) = argmin_{x:valid(x)|\pi - x|}$ ensures the updated policy is a valid probability distribution.

Algorithm 1 summarizes the WPL learning procedure for the learning agent.

Algorithm 1 Simplified Weighted Policy Learning (WPL) Algorithm for the learning agent

1: Initialize: Learning rate θ, discount factor ζ, and policy learning rate η.
2: Set all $Q(s, a)$ to 0 and policy $\pi(s)$ to initial values (ICs).
3: while the task is not finished do:
4: Choose action at in state st using $\pi_t(s_t)$ with exploration.
5: Receive reward r_t and next state s_{t+1}.
6: Update $Q(s_t, a_t)$ using Eq. (1).
7: Compute $V(s_t)$.
8: for each action a do:
9: Update $\pi_{t+1}(s_t, a)$ using Eq. (2).
10: end for
11: Normalize policy: $\pi_{t+1}(s_t) = limit\big(\pi_{t+1}(s_t)\big)$
12: end while

Tow-Tank Liquid Level System

A. System Description:

The two-tank liquid level system consists of two vertically placed tanks. Tank 1 receives input from the exterior and output to Tank 2. Tank 2's output goes to the environment. The two tanks should be cylindrical with a constant cross-sectional area and incompressible fluid.

The following notations can be employed:

- A_1, A_2: Cross-sectional area of Tank 1 and Tank 2 [m^2].
- $h_1(t)$, $h_2(t)$: Liquid heights in Tank 1 and Tank 2 at time t [m].
- $q_{in}(t)$: Rate of inflow into Tank 1 [m^3/s].
- $q_{12}(t)$: Rate of flow from Tank 1 into Tank 2 [m^3/s].
- $q_{out}(t)$: Rate of outflow from Tank 2 to the environment [m^3/s].
- C_1, C_2: Flow coefficients for Tank 1 and Tank 2 discharge outlets.

Assuming flow rates follow Torricelli's Law for free outflow under gravity [2], the inter-tank and output flow rates are given by:

$$q_{12}(t) = C_1\sqrt{h_1(t) - h_2(t)} \qquad (3)$$

$$q_{out}(t) = C_2\sqrt{h_2(t)} \qquad (4)$$

B. Continuous-Time Model

Imposing a mass balance on the two tanks results in the following first-order nonlinear differential equations [3]:

Tank 1:

$$A_1 \frac{dh_1(t)}{dt} = q_{in}(t) - C_1\sqrt{h_1(t) - h_2(t)} \qquad (5)$$

Tank 2:

$$A_2 \frac{dh_2(t)}{dt} = C_1\sqrt{h_1(t) - h_2(t)} - C_2\sqrt{h_2(t)} \qquad (6)$$

These equations capture the nonlinear behavior of the tank system due to the square-root flow relationships.

C. Discrete-Time Model:

Using the forward Euler method for discretization with a sampling time T_s, the discrete-time model becomes [2], [3]:

$$h_1(k+1) = h_1(k) + \frac{T_s}{A_1}\left(q_{in}(k) - C_1\sqrt{h_1(k) - h_2(k)}\right) \qquad (7)$$

$$h_2(k+1) = h_2(k) + \frac{T_s}{A_2}\left(C_1\sqrt{h_1(k) - h_2(k)} - C_2\sqrt{h_2(k)}\right) \qquad (8)$$

This discrete model is suitable for digital control design and simulation applications.

Control Framework and Reward Design

A. State Discretization Method:

To enable the tabular learning of the WPL, continuous system states are discretized into a finite set of indices. Specifically, the state is defined by the error between the setpoint and the level of Tank 2, $h_{error} = h_{ref} - h$. The error is first limited to the range $[h_{error_{min}}, h_{error_{max}}]$ in order to limit the impact of outliers. A non-homogeneous binning technique is subsequently applied to both variables. The space of errors is quantized more finely near zero (i.e., near the setpoint) to enhance accuracy of control in sensitive areas.

B. State-Action Space:

Every continuous variable is mapped onto the closest bin, and the generated bin indices are concatenated into a single discrete state index through row-major flattening. This creates a compact and resolution-aware discrete state representation that allows efficient learning and generalization in reinforcement learning-based control tasks. Actions a_t are chosen through an ϵ-greedy policy to trade-off exploration and exploitation. The actions are specified as follows in Table 1:

Table 1: The actions of the learning agent.

Actions	Values
0	0.0
1	0.2
2	0.4
3	0.8
4	1.2
5	1.6
6	2.2

C. Reward Function for the WPL algorithm in a Two-Tank Level System:

The objective of the learning here is to maintain the level of the second tank, $h_2(t)$, as close as possible to the level setpoint h_{ref}. The WPL algorithm selects the proper action at each state that will lead eventually to minimize the level error.

The reward function at time t is formulated as:

$$r(t) = -\alpha_1 e(t)^2 - \alpha_2 \Delta e(t)^2 - \alpha_3 \Delta u(t)^2 \qquad (9)$$

where, α_1, α_2, $\alpha_3 > 0$ are weighting coefficients,

$$e(t) = h_{\text{ref}} - h_2(t)$$

is the error at time$t$,

$$\Delta e(t) = e(t) - e(t-1)$$

is the change in error, and

$$\Delta u(t) = u(t) - u(t-1)$$

is the change in controller output.

- The term $-\alpha_1 e(t)^2$ penalizes large deviations from the reference level, encouraging accuracy.
- The term $-\alpha_2 \Delta e(t)^2$ encourages stability, discourages oscillations, and penalizes rapid changes in error.
- The term $-\alpha_3 \Delta u(t)^2$ promotes smooth control signals and penalizes aggressive control actions.

Simulation and Results

A. Simulation Setup:

The WPL is implemented on a two-tank liquid level system in order to regulate the level of the second tank, h_2, to a target reference level h_{ref} .

The dynamics of the two-tank liquid level system are defined as follows:

$$Q_{\text{out1}} = 0.3\sqrt{h_1}$$

$$Q_{\text{out2}} = 0.3\sqrt{h_2}$$

h_1 and h_2 represent the fluid levels in tank 1 and tank 2, respectively. The cross-sectional areas of both tanks and the timestep used are defined as A_1 = 1.0, A_2 = 0.8, Δ_t = 0.05 seconds. During the learning phase, the setpoint level h_{ref} was randomly initialized within the range 5 to 10 meters. The reinforcement learning agent uses discretized states based on the level error ($error = h_{ref} - h$). The level error is clamped within the range [−1, 1] to avoid excessively large deviations and linearly spaced in a total of 107 states:

- 33 linearly spaced bins in [−1,−0.2],
- 41 linearly spaced bins in [−0.1, 0.1] (high-resolution region near the setpoint),
- 33 linearly spaced bins in [0.2, 1].

B. The WPL Algorithm Parameters:

The WPL algorithm parameters were set as follows:

- Number of episodes: 50.
- Discount factor: ζ= 0.80.
- Learning rate start: θ = 0.7, decaying to a minimum of 0.1.
- Exploration rate (epsilon) start: ϵ = 1.0, decaying to a minimum of 0.1.
- Policy update rate start: η = 0.1.
- Maximum episode length: 2000 time steps.

C. Results:

The WPL's learning agent was trained over 50 episodes. During testing, the reference level was set to h_{ref} = 10.0 meters initially. The system was simulated for 10,000 steps with reference level changes at t = 3000 (set to 5.0 m) and t = 6000 (back to 10.0 m) to test tracking performance. Figure 1 shows the level of the second tank over time. The figure shows that the WPL algorithm were successfully able to regulate the level of the second tank to the reference with fast convergence and minimal overshoot, despite sudden changes in the target level. The figure also shows that the response obtained by the WPL algorithm is better than the response obtained by the PID controller in terms of fast convergence and lower overshoot.

D. Discussion:

The performance of the WPL algorithm in a two-tank liquid level system was evaluated in the simulation experiments against the performance of a conventional PID controller. The two-tank liquid level system is a nonlinear system because it's governed by the nonlinear flow equations. In the simulation experiments, the objective was to regulate the level of the second tank (h_2) so that it tracks a reference setpoint (h_{ref}), where the setpoint is changed at t = 3000s (10m → 5m) and at t = 6000s (5m → 10m). The simulation results show that the WPL algorithm succeeded to maintain the level of the second tank at the reference setpoint, with fast convergence and minimal overshoot, despite sudden changes in the target level. The simulation results also show that the WPL algorithm outperformed the conventional PID controller in terms of fast convergence and lower overshoot.

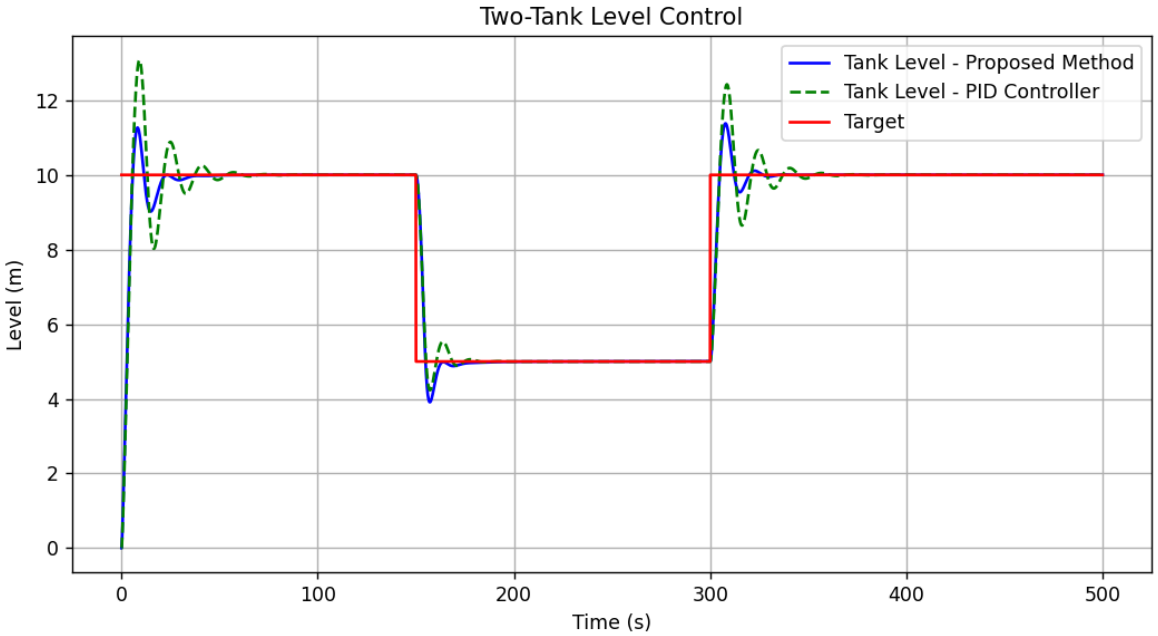

Figure 1: Tank level h_2 response over time under WPL control with varying reference levels h_{ref}.

Conclusion

In this paper, a new application of the Weighted Policy Learning (WPL) algorithm to the nonlinear two-tank liquid level control problem is proposed. In contrast to conventional model-based controllers like PID or MPC that need proper modeling and tuning, the WPL algorithm learns policies by experience without knowing system dynamics. The WPL algorithm was also able to effectively deal with the coupled dynamics, between the two tanks, and learn non-stationary behavior. To allow tabular learning, system states were discretized according to a non-uniform binning policy that gave higher priority to accuracy near the setpoint. By framing the adjustment of the second tank's level as a reinforcement learning problem, the WPL algorithm succeeded to regulate the level at its desired setpoint, despite of the system's nonlinearity and the sudden changes in the desired level target. Simulation experiments demonstrated significant improvement over conventional PID control in terms of fast convergence, low overshoot, and zero steady-state error, without human tuning.

References

[1] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction," 2nd ed. Cambridge, MA: MIT Press, 2018.

[2] K. Ogata, "Modern control engineering," 5th ed. Upper Saddle River, NJ: Prentice Hall, 2010.

[3] K. J. Åström and R. M. Murray, "Feedback systems: An introduction for scientists and engineers," Princeton, NJ: Princeton University Press, 2010.

[4] S. Abdallah and V. Lesser, "A multiagent reinforcement learning algorithm with non-linear dynamics," Journal of Artificial Intelligence Research 33, pp. 521-549, 2008.

[5] H. M. Schwartz, "Multi-agent machine learning: A reinforcement approach," John Wiley and Sons, Inc., New York, 2014.

[6] W. Hinojosa, S. Nefti and U. Kaymak, “Systems control with generalized probabilistic fuzzy-reinforcement learning,” Fuzzy Systems, IEEE Transactions on, 19.1, pp. 51-64, 2011.
[7] M. D. Awheda and H. M. Schwartz, “The residual gradient FACL algorithm for differential games,” Electrical and Computer Engineering (CCECE), IEEE 28th Canadian Conference on, pp. 1006-1011, 2015.
[8] M. D. Awheda and H. M. Schwartz, “A Residual Gradient Fuzzy Reinforcement Learning Algorithm for Differential Games,” International Journal of Fuzzy Systems, Vol. 19, No. 4, pp. 1058-1076, 2017.
[9] W. Dixon, “Optimal adaptive control and differential games by reinforcement learning principles,” Journal of Guidance, Control, and Dynamics, 37.3, pp. 1048-1049, American Institute of Aeronautics and Astronautics, 2014.
[10] S. Sen and G. Weiss, “Learning in multiagent systems, In: Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence,” The MIT Press, Cambridge, 1999.
[11] A. Dev, K. R. Chowdhury and M. P. Schoen, “Q-Learning Based Control for Swing-Up and Balancing of Inverted Pendulum,” 2024 Intermountain Engineering, Technology and Computing (IETC), Logan, UT, USA, pp. 209-214, 2024.
[12] G. Puriel, Li, and B. Ovilla-Martinez “Robot PID control using reinforcement learning,” 2023 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 885-890, 2023.
[13] P. Balazy, K. Pieprzycki and P. Knap, “Robust reinforcement learning for overhead crane control with variable load conditions,” 25th International Carpathian Control Conference (ICCC), Poland, pp. 01-06, 2024.
[14] D. Lee, S. Jin and C. Lee, “Deep reinforcement learning of semi-active suspension controller for vehicle ride comfort,” in IEEE Transactions on Vehicular Technology, vol. 72, no. 1, pp. 327-339, 2023.
[15] F. A. Yaghmaie, F. Gustafsson and L. Ljung, “Linear quadratic control using model-free reinforcement learning,” in IEEE Transactions on Automatic Control, vol. 68, no. 2, pp. 737-752, Feb. 2023.
[16] D. S. Bhandare, V. S. Jape, H. H. Kulkarni, S. M. Mahajan, P. Sable and Y. Pawar, “Automatic liquid level control of two tank system using PLC,” International Conference on Intelligent Systems and Advanced Applications (ICISAA), Pune, India, pp. 1-6, 2024.
[17] A. Baciu and C. Lazar, “Parameters setting of a data-driven model-free adaptive controller for a coupled two-tank system,” 26th International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, Romania, pp. 397-402, 2022.