A Comparative Study of Reinforcement Learning and Analytical Methods for Optimal Control

1st Myeongseok Ryu
School of Mechanical Engineering
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
dding_98@gm.gist.ac.kr

2nd Junseo Ha
Artificial Intelligence Graduate School
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
hajunseo@gm.gist.ac.kr

3rd Minji Kim
School of Mechanical Engineering
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
kalswl326@gm.gist.ac.kr

4th Kyunghwan Choi*
School of Mechanical Engineering
Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
khchoi@gist.ac.kr

2023 International Workshop on Intelligent Systems (IWIS) | 979-8-3503-0504-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/IWIS58789.2023.10284677

***Abstract*—Numerous reinforcement learning (RL) algorithms have been introduced to resolve challenging tasks like game playing, natural language processing, and control. Particularly, RL can be used to find a good policy for control systems for which the optimal control sequence is difficult to find by analytical methods. This paper compares RL and analytical methods for optimal control in an inverted pendulum environment. Dynamic programming (DP) and model predictive control (MPC) are considered for the analytical methods. The control results of RL, DP, and MPC are qualitatively and quantitatively compared in terms of total reward, state response, and control sequence to investigate the relationships between them. Because they have similar problem formulations, the relationships can be explained by RL parameters: discounting factor and exploration rate. This comparative study is expected to provide insights to those studying RL algorithms and optimal control theories.**

***Index Terms*—Optimal Control, Reinforcement Learning, Dynamic Programming, Model Predictive Control**

I. Introduction

Recently, there has been considerable research interest in reinforcement learning (RL) to address challenging control problems characterized by complex dynamics and multiple constraints. This is because RL uses a model-free approach; a policy (i.e., control law) is trained based on experience obtained through interactions with the environment. The analytical model is not explicitly used in the training process but is contained in the environment. Using the model-free approach differentiates the RL from analytical optimal control methods that explicitly use the analytical model to solve optimal control problems.

Nonetheless, RL shares some fundamental concepts and principles with analytical optimal control methods (e.g., dynamics programming (DP) and model predictive control (MPC)), because RL's origins can be traced back to the principles of optimal control theory. Several prior studies have explored the relationships between RL and the analytical optimal control methods. In [1], DP's cost function and control law is compared with the value function and optimal policy of Q-learning. In [2], [3], MPC and Deep Q-Networks (DQN) are compared in terms of the cost function structure, constraint satisfaction, and trade-off between performance and cost. However, as per the authors' knowledge, a comprehensive comparative study encompassing RL, DP, and MPC has not been conducted yet.

This paper investigates the relationships between RL and the two analytical optimal control methods, DP and MPC, through a comprehensive comparison of their control results based on total cost (or reward), state response, and control sequence. The comparative study is conducted within the widely recognized inverted pendulum environment. The relationships between methods are primarily explained through RL parameters: the discounting factors and exploration rate.

II. Environment

A. Inverted Pendulum Environment

The inverted pendulum problem is a classic control problem, widely recognized in control theory and reinforcement learning. The main objective of this environment is to achieve the optimal regulation of the pendulum by swinging it to the upside.

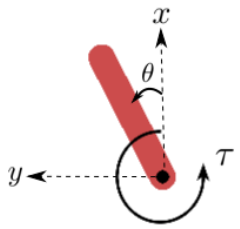

Fig. 1. Inverted Pendulum Environment.

The environment is depicted in Fig. 1. The system dynamics are as follows:

$$\begin{bmatrix}\theta_{k+1}\\ \dot{\theta}_{k+1}\end{bmatrix} = \begin{bmatrix}\theta_k + \dot{\theta}_k T_s \\ \dot{\theta}_k + \left(\frac{3g}{2l}\sin(\theta_k) + \frac{3\tau}{ml^2}\right)T_s\end{bmatrix} = h(\theta_k, \dot{\theta}_k, \tau), \tag{1}$$

where θ is the angle, $\dot{\theta}$ is the angular velocity, T_s is the sampling time, g is the gravitational acceleration, l is the rod length, τ is the torque, and m is the rod mass.

The control problem of swinging up the pendulum can be formulated as an optimal control problem, outlined as follows:

$$\begin{aligned}&\text{Find } u_k,\ k \in [0, N-1]\\ &\text{to minimize } J = \sum_{k=0}^{N-1} l_k = \sum_{k=0}^{N-1} \|\mathbf{x}_k\|_{\mathbf{Q}}^2 + \|u_k\|_R^2\\ &\text{subject to}\\ &\mathbf{x}_{k+1} = h(\mathbf{x}_k, u_k)\\ &|\dot{\theta}| \le \dot{\theta}_{\max}\\ &|u| \le u_{\max}\end{aligned} \tag{2}$$

where $\mathbf{x} = [\theta, \dot{\theta}]^T$ is the state, $u = \tau$ is the control input (or action), $\mathbf{Q}$ is a diagonal matrix with the entries 1 and 0.1 and scalar R is 0.001 (i.e., the cost function is $l_k = \theta_k^2 + 0.1\dot{\theta}_k^2 + 0.001\tau_k^2$). The parameters used to define this problem are selected the same as in [4] and listed in Table I.

Despite the cost function being represented in a convex quadratic form, the problem remains a nonconvex optimization challenge due to the presence of nonlinear system dynamics. In addition, the constraint on the control input may prevent the pendulum from reaching the target position (i.e., $\theta = 0$) with one swing; the anticipated strategy involves utilizing multiple swings to achieve sufficient kinetic energy. Therefore, this problem is challenging to solve using analytical optimal control methods.

III. Theoretical Background

A. Analytical Optimal Control Methods

The analytical optimal control methods can be used to find the optimal or suboptimal control law if the system model is known. Among them, DP [5] guarantees to provide the global solution to optimal control problems over a time horizon, such as (2). Nonetheless, DP is not well-suited for real-time control law because of its challenging implementation and the substantial computation resources it demands.

MPC [6] is considered a more practical method compared to DP, since it addresses more compact optimal control problems.

TABLE I
System and Control Problem Parameters

Parameter	Value	Parameter	Value
m (kg)	1	l (m)	1
g (m/s^2)	10	T_s (s)	0.05
$\dot{\theta}_{\max}$ (rad/s)	8	$u_{\max}$ (s)	2

At each time step, MPC formulates an optimal control problem over a receding prediction horizon, which is typically smaller than the time horizon considered in DP. Then, MPC resolves this optimization problem to obtain the optimal sequence of control inputs and state response.

B. RL

In the RL field, numerous methods are introduced to exceed the performances of conventional controllers. Because RL approximates its policy function, which corresponds to the control law in the control theory, to an optimal policy by trial and error method, it does not require the analytical system model.

In particular, DQN [7] has garnered considerable attention from researchers as a fundamental but powerful RL method. The DQN is an extension of the Q-learning algorithm that uses deep learning networks to estimate and approximate its action-value function (Q-function). The Q-function is updated by the following rule:

$$\begin{aligned}Q(\mathbf{x}_k, u_k) \leftarrow & Q(\mathbf{x}_k, u_k) + \alpha r(\mathbf{x}_k, u_k)\\ &+ \alpha(\gamma \max Q(\mathbf{x}_{k+1}, \cdot) - Q(\mathbf{x}_k, u_k)),\end{aligned} \tag{3}$$

where r is the reward earned at each step, γ is the discounting factor, and α is the learning rate.

Two representative hyperparameters of DQN are the discounting factor γ and the exploration rate ϵ. Using the discounting factor ranging from 0 to 1, the total reward that the agent has earned in an episode is discounted as follows:

$$G_k = \sum_{k=0}^{N-1} \gamma^{k-1} r_k(\mathbf{x}, u) \tag{4}$$

A low value of the discounting factor causes the agent to be attracted to instant rewards, and a high value makes future rewards more attractive. Additionally, selecting a discounting factor smaller than 1 ensures the total reward is finite.

The exploration rate makes the agent explore the entire environment to evaluate its Q-function through random actions.

C. Relationship between Optimal Control Theory and RL

In the optimal control theory and RL, the targets to minimize and maximize are the total cost and total reward, respectively. If the discounting factor is 1, the total reward of RL is in the same form as the total cost of the optimal control problem as follows:

$$G_k = \sum_{k=0}^{N-1} r_k, J = \sum_{k=0}^{N-1} l_k. \tag{5}$$

This means that the DQN with a discounting factor close to 1 can approximate its policy to the global optimal policy obtained by DP.

On the other hand, using a low gamma can make the RL agent short-sighted, which is similar to using a short prediction horizon in MPC. This is because the discounting factor reduces the values of future rewards as the far future is not considered in MPC with a short prediction horizon.

IV. EXPERIMENT

The experiment was designed to investigate the relationships mentioned in Section III-C. The inverted pendulum, initially located at $\mathbf{x}_0 = [3, 0]^T$, was controlled by two analytic optimal control methods, DP and MPC, and RL. DQN algorithm was used to implement RL. The MPC used the prediction horizon of 50, which was the maximum value for reasonable computing time, and utilized the linearized system at its current state instead of its original nonlinear dynamics. The hyperparameters of DQN were selected as in Table II. In addition, various combinations of the discounting factor and the initial exploration rate were examined for DQN The exploration rate decayed from each initial value to 0.2 during the first 250 episodes. The action space of DQN was discretized into 41 actions. The performance of DP, MPC, and DQN was evaluated based on the state response, control sequence, and total cost.

A. Analytical Optimal Control Methods

Figure 2 illustrates the results of the DP and MPC controllers. The DP solution shows the optimal control sequence that efficiently swings up the pendulum counter-clockwise in approximately 60 steps. However, the MPC controller attempts to swing the pendulum clockwise, even though the torque is not sufficient to make a move in clockwise. This behavior is attributed to the limited prediction horizon for feedback control, preventing the controller from discovering a control solution that efficiently utilizes gravity. The MPC controller is expected to require more than 50 steps to achieve a control sequence like the DP controller.

TABLE II
COMMON DQN HYPERPARAMETERS

Hyperparameter	Value
Train Episode	1000
Batch Size	128
Replay Buffer Size	10000
Learning Rate	1e-4
Hidden Layer Number	1
Nodes per Layer	128

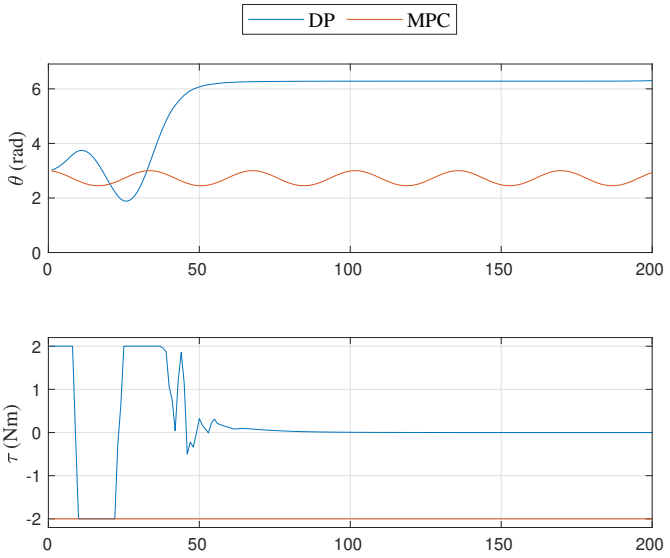

Fig. 2. Comparison of DP and MPC.

B. DQN with Various Discounting Factors

To examine the effect of the discounting factor, DQN agents are trained with discounting factors of 1, 0.75, and 0.5. The exploration rate of all agents, which decays to 0.2 during 250 episodes has an initial value of 0.99. As shown in Fig. 3, the agent with a discounting factor 1 showed the control result that is the closest to the optimal result obtained by DP. Conversely, agents with discounting factors of 0.5 and 0.75 exhibited behavior similar to the control sequence obtained by MPC. This observation aligns with the explanation in Section III-C. The agent's preference for immediate rewards leads to swinging up the pendulum in a clockwise direction. A similar result was reported in [8].

The control results obtained using discounting factors of 1, 0.95, and 0.9 are shown in Fig. 4. The lower the discounting factor is, the slower the state response is. This is because immediate rewards are more appreciated when using a low discounting factor.

C. DQN with Various Exploration Rates

Figure 5 shows the control results obtained with various initial values of the exploration rate. Notably, the agent with zero exploration rate, which exclusively took greedy actions without exploration, achieved the control result closest to DP. It is because the initial state space of the pendulum coincides with the entire state space. Consequently, even though the agent took greedy actions, it was possible to estimate informative state and action values accurately and thus approximate the Q-value function close to the optimal policy.

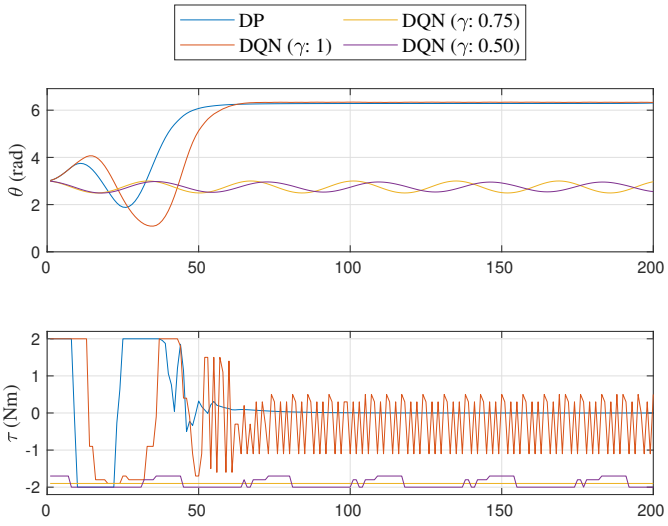

Fig. 3. Comparison of DP and DQN with discounting factors of 1, 0.75, and 0.5.

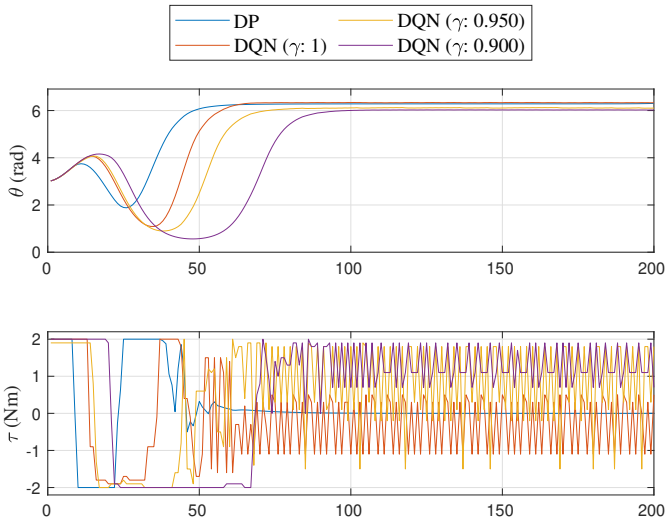

Fig. 4. Comparison of DP and DQN with discounting factors of 1, 0.95, and 0.9.

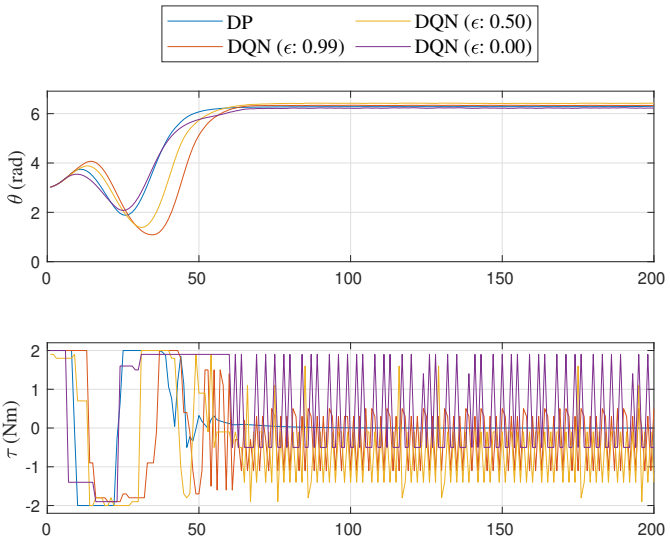

Fig. 5. Comparison of DP and DQN with various initial values of the exploration rate.

TABLE III
COMPARISON OF TOTAL COSTS

Case Name	Total Cost	Case Name	Total Cost
DP	314.167	MPC	1494.016
DQN (γ: 1.00)	345.242	DQN (ϵ: 0.99)	345.242
DQN (γ: 0.75)	1518.782	DQN (ϵ: 0.50)	339.165
DQN (γ: 0.50)	1521.538	DQN (ϵ: 0.00)	352.418
DQN (γ: 0.95)	360.987	DQN (γ: 0.90)	386.856

D. Comparison of Total Costs

Table III shows the total cost of each control scheme. The global optimal solution obtained by DP achieved the lowest total cost. The DQN agent with a discounting factor of 1 and zero exploration rate achieved the second lowest total cost. This DQN agent is expected to have a total cost closer to that of DP when the Q-function is more accurately approximated to the optimal policy with a deeper network. On the other hand, cases where the pendulum cannot be successfully swung up, such as MPC and DQN cases with low discounting factors, demonstrated total costs exceeding 1000.

E. Differences in Control Sequences

As can be seen in Figs. 3, 4, and 5, the control sequences of all DQN agents exhibited a bang-bang control law pattern once reaching the target position ($\theta = 0$). One of the main reasons for this is that the changing rate of actions was not considered in the training. The bang-bang pattern is expected to be removed by incorporating a penalty term for the changing rate of actions in the reward function. By contrast, DP did not show the bang-bang pattern even though the charging rate of control inputs was not considered. This is because DP solved the optimal control problem analytically; thus, reasonable behavior was obtained.

While the bang-bang control law did not directly affect the total reward in the RL training process, it could be an issue when applying this policy to the real-world system. High-frequency control inputs are often undesired for hardware controllers, as they can cause stress and inefficiency. Regulating the changing rate of actions is essential for ensuring safe and efficient control in real-world applications

V. CONCLUSION

This paper investigated the relationships between the analytical optimal control methods, DP and MPC, and RL. The similarity of RL to DP and MPC was explained based on how much future rewards were appreciated using the discounting factor, a hyperparameter of RL. The effect of another hyperparameter of RL, the exploration rate, was also investigated. The experiment demonstrated that 1) the DQN agent with a discounting factor of 1 showed a control result close to the optimal control sequence obtained by DP, and 2) the DQN agent with a lower discounting factor showed a similar control result to MPC. An interesting finding was that the DQN agent finds a better policy without exploration than with a nonzero exploration rate because the initial state space of the pendulum coincides with the entire state space.

Understanding the relationship between the analytic optimal control methods and RL is crucial for unifying the two approaches for optimal control. Future work concerns deeper analysis of the relationship between the two approaches and developing integrated methodologies that leverage the advantages of each approach.

REFERENCES

[1] R. Sutton, A. Barto, and R. Williams, "Reinforcement learning is direct adaptive optimal control," *IEEE Control Systems Magazine*, vol. 12, no. 2, pp. 19–22, 1992.

[2] J. Arroyo, C. Manna, F. Spiessens, and L. Helsen, "Reinforced model predictive control (RL-MPC) for building energy management," *Applied Energy*, vol. 309, p. 118346, 2022.

[3] D. Görges, "Relations between model predictive control and reinforcement learning," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 4920–4928, 2017.
[4] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.
[5] D. Kirk, *Optimal Control Theory: An Introduction*. Dover Books on Electrical Engineering Series, Dover Publications, 2004.
[6] L. Wang, *Model Predictive Control System Design and Implementation Using MATLAB*. Springer Publishing Company, Incorporated, 1st ed., 2009.
[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.
[8] Y. Fu, S. Xu, Q. Zhu, Z. O'Neill, and V. Adetola, "How good are learning-based control vs model-based control for load shifting? investigations on a single zone building energy system," *Energy*, vol. 273, p. 127073, 2023.