Điều khiển xe hai bánh tự cân bằng mô hình bất định dựa trên phương pháp quy hoạch động thích nghi

Conference Paper · July 2019

DOI: 10.15625/vap.2019000270

CITATION
1

READS
842

1 author:

Nam Nguyễn
Hanoi University of Science and Technology
22 PUBLICATIONS **43** CITATIONS

SEE PROFILE

Tuyển tập Hội nghị khoa học toàn quốc lần thứ nhất về Động lực học và Điều khiển
Đà Nẵng, ngày 19-20/7/2019, tr. 140-146, DOI 10.15625/vap.2019000270

Điều khiển xe hai bánh tự cân bằng mô hình bất định dựa trên phương pháp quy hoạch động thích nghi

Trần Gia Khánh, Lê Việt Anh, Phan Anh Tuấn, Phan Xuân Minh và Nguyễn Hoài Nam
Bộ môn Điều khiển tự động, Viện Điện, Đại học Bách khoa Hà Nội
Số 1 Đại Cồ Việt, Quận Hai Bà Trưng, Hà Nội, Việt Nam
E-mail: nam.nguyenhoai@hust.edu.vn

Tóm tắt
Bài báo ứng dụng phương pháp quy hoạch động thích nghi sử dụng cấu trúc Actor-Critic cho xe hai bánh tự cân bằng (XHBTCB). Việc sử dụng phương pháp quy hoạch động thích nghi đã giảm thiểu đáng kể công sức và thời gian giải bài toán điều khiển tối ưu, khi không phải giải trực tiếp bằng các phương pháp giải tích và cũng không cần thiết xây dựng mô hình toán học đầy đủ của xe. Ngoài ra, bộ điều khiển tối ưu cũng sẽ tự cập nhật để đáp ứng với thay đổi của hệ thống, do thuật toán điều khiển chỉ sử dụng các biến trạng thái phản hồi đo được. Mô phỏng số trên phần mềm MATLAB được tiến hành để đánh giá chất lượng của thuật toán điều khiển.

Từ khóa: Điều khiển tối ưu thích nghi, Quy hoạch động thích nghi, Xe hai bánh tự cân bằng.

1. Giới thiệu

Điều khiển tối ưu là một trong những lĩnh vực nhận được nhiều sự quan tâm của các nhà nghiên cứu về lý thuyết điều khiển hiện đại. Luật điều khiển tối ưu thiết kế không chỉ ổn định hệ thống mà còn tối thiểu hàm chi phí mô tả chỉ tiêu chất lượng mong muốn. Lời giải cho bài toán điều khiển tối ưu có thể thu được bằng việc sử dụng nguyên lý cực đại của Pontryagin hoặc tìm nghiệm của phương trình HJB. Cả hai cách tiếp cận trên đều có nhược điểm chung là yêu cầu thông tin đầy đủ về hệ thống, bao gồm các biến trạng thái và mô hình động học. Trong trường hợp mô hình hệ thống chỉ là gần đúng hoặc có yếu tố bất định thì bộ điều khiển tối ưu thu được bằng phương pháp giải tích hoặc phương pháp số có thể không mang lại hiệu quả điều khiển tối ưu khi áp dụng lên hệ thống thực. Trong khi đó, điều khiển thích nghi được phát triển để giải quyết các bài toán điều khiển với mô hình bất định hoặc khó xác định đủ chính xác. Phương pháp thích nghi thường tập trung vào thiết kế luật điều khiển không sử dụng các yếu tố bất định, hoặc xấp xỉ các yếu tố bất định sao cho vẫn đảm bảo hiệu quả của hệ thống kín, không nhất thiết phải đảm bảo tối ưu theo một nghĩa nào đó. Kết hợp các ưu điểm của điều khiển tối ưu và điều khiển thích nghi, điều khiển tối ưu thích nghi được phát triển bằng cách bổ sung yếu tố tối ưu trong thiết kế điều khiển thích nghi, ví dụ như thông số bộ điều khiển là một biến của bài toán tối ưu hóa, hoặc bổ sung yếu tố thích nghi trong thiết kế điều khiển tối ưu, ví dụ như xấp xỉ các thông số hệ thống được sử dụng trong luật điều khiển tối ưu. Xem xét một ví dụ của bài toán điều khiển tối ưu thích nghi như sau. Thông thường, một bài toán điều khiển tối ưu sẽ được giải quyết nếu phương trình HJB được giải. Đối với hệ tuyến tính, phương trình HJB trở thành phương trình đại số Riccati (Algebraic Riccati Equation - ARE). Nếu ma trận trạng thái (A, B) của hệ tuyến tính có sẵn, nghiệm ARE hoàn toàn có thể tìm được bằng giải tích. Ngược lại, nếu thiếu một trong các ma trận này thì phương pháp giải tích không thể áp dụng. Đối với hệ phi tuyến, phương trình HJB trở thành phương trình vi phân phi tuyến. Nghiệm giải tích của phương trình HJB phi tuyến thậm chí nói chung là không thể giải ngay cả với hệ thống có mô hình xác định. Để khắc phục hạn chế nêu trên, nhiều giải thuật xấp xỉ nghiệm của phương trình ARE hoặc HJB dựa trên lý thuyết cơ sở của học tăng cường (Reinforcement Learning) đã được đề xuất.

Một bài toán học tăng cường thường xem xét một cá thể (agent) có tương tác với môi trường bên ngoài bằng một chuỗi các hành động (actions) và nhận được các thành quả (reward), có thể là một chỉ tiêu chất lượng đại diện bằng một hàm chi phí (cost), từ môi trường. Phương pháp học tăng cường là một nhánh của học máy (Machine Learning), nhằm thu được chính sách (policy), chính sách này có thể hiểu là một quá trình hoạt động hay luật điều khiển, tối ưu cho một cá thể dựa trên các đáp ứng quan sát được từ tương tác giữa cá thể và môi trường [1]. Một thuật toán học tăng cường nói chung có hai bước, đầu tiên mỗi cá thể đánh giá thành quả của một chính sách hiện tại thông qua tương tác với môi trường, bước này được gọi là Đánh giá chính sách (Policy Evaluation). Tiếp theo dựa trên thành quả đã đánh giá, cá thể tiến hành cập nhật chính sách nhằm tăng chất lượng, tương đương với tối thiểu hóa hàm chi phí. Bước này được đặt tên là Cải tiến chính sách (Policy Improvement). Thời gian gần đây, các nhà nghiên cứu đang tập trung vào hướng áp dụng kỹ thuật học tăng cường trong điều khiển phản hồi các hệ thống động học. Một trong các phương pháp phổ biến của học tăng cường được ứng dụng trong điều khiển là kỹ thuật lặp PI (Policy Iteration) [2]. Thay vì sử dụng các phương pháp toán học để giải trực tiếp phương trình HJB, thuật toán PI bắt đầu bằng việc đánh giá hàm chi phí của một luật điều khiển khởi tạo chấp nhận được (admissible control policy). Công việc này thường thu được bằng việc giải phương trình Lyapunov phi tuyến [3]. Hàm chi phí mới này được sử dụng để cải tiến luật điều khiển, tương đương với tối thiểu hóa hàm Hamilton ứng với hàm chi phí đó. Quá trình lặp hai bước này được tiến hành cho tới

khi luật điều khiển hội tụ tới luật điều khiển tối ưu.

Với sự phát triển của học tăng cường, nhiều phương pháp thời gian thực đã được áp dụng để tìm luật điều khiển tối ưu trực tuyến mà không cần hiểu biết hoàn toàn chính xác về động lực học của hệ thống, cách tiếp cận này thường được gọi là quy hoạch động thích nghi (Adaptive Dynamic Programming - ADP) [4], trong nhiều tài liệu cũng được gọi là quy hoạch động xấp xỉ (Approximate Dynamic Programming) [1]. Dựa trên khả năng có thể xấp xỉ hàm phi tuyến trơn, mạng nơron thường được sử dụng cho việc thực thi các thuật toán học lặp. Các thuật toán sẽ được thực thi trực tuyến trên cấu trúc Actor-Critic, bao gồm hai mạng nơron xấp xỉ hàm, mạng thứ nhất được gọi là Actor, dùng để xấp xỉ luật điều khiển, mạng thứ hai được gọi là Critic đại diện cho hàm chi phí. Đối với hệ tuyến tính liên tục, nghiên cứu [5] đã giới thiệu hai thuật toán lặp PI ngoại tuyến, tương đương về mặt toán học với phương pháp Newton. Các phương pháp này đã loại bỏ được yêu cầu về mô hình nội động học của hệ thống (mô hình không xét tới kích thích bên ngoài) bằng việc đánh giá hàm chi phí ứng với luật điều khiển trên một quỹ đạo trạng thái ổn định, hoặc bằng sử dụng biến trạng thái đo được để xây dựng phương trình Lyapunov. Phát triển hướng nghiên cứu của Murray, trong [6], Vrabie và các cộng sự trình bày thiết kế điều khiển sử dụng học tăng cường để giải trực tuyến bài toán điều khiển tối ưu tuyến tính toàn phương (Linear Quadratic Regulator - LQR). Cụ thể, phương pháp sử dụng thuật toán lặp PI dựa trên dữ liệu động học đo được để giải lặp phương trình Riccati. Trong thiết kế, ma trận nội động học của hệ thống cũng được loại bỏ trong quá trình thiết kế, nhưng ma trận ngoại đông học (mô tả quan hệ giữa tác động bên ngoài đối với trạng thái hệ thống) vẫn cần sử dụng, do đó còn gọi là thuật toán cho hệ bất định một phần (partially model-free). Phương pháp cho hệ bất định hoàn toàn (fully model-free) được phát triển trong [7], với việc sử dụng tín hiệu nhiễu thăm dò thêm vào tín hiệu đầu vào trong quá trình học. Đối với hệ phi tuyến, trong [8] và [9], thuật toán trực tuyến cho hệ phi tuyến dạng affine bất định một phần được trình bày, mang tới lời giải cục bộ cho phương trình HJB phi tuyến. Phương pháp cho hệ bất định hoàn toàn được trình bày trong công trình [4], có thể coi là mở rộng cho phương pháp của hệ tuyến tính trong [7]. Tuy chỉ là phương pháp tối ưu ổn định bán toàn cục (semi-global), do chưa đảm bảo sự ổn định hoàn toàn mà chỉ trong trường hợp thỏa mãn các giả thiết nhất định, nhưng cũng đã là một bước đột phá khi có thể tìm ra luật điều khiển tối ưu mà có thể loại bỏ hoàn toàn yêu cầu về mô hình của hệ thống. Mở rộng kết quả, các tác giả đã trình bày phương pháp ổn định toàn cục cho một lớp hệ đa thức (các hàm động học có dạng đa thức) ở trong [10].

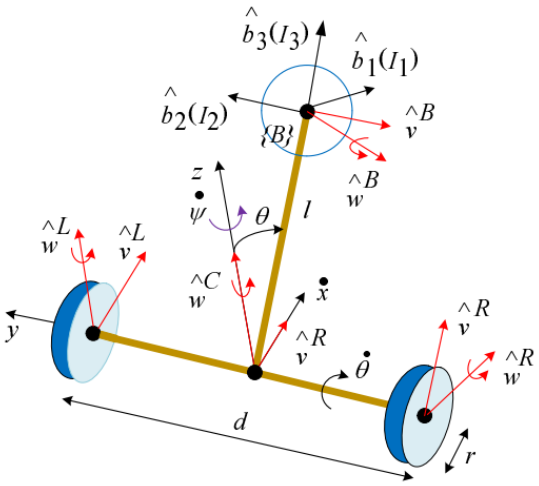
Như vậy, có thể thấy bằng việc áp dụng học tăng cường và quy hoạch động thích nghi, không những bài toán tối ưu được giải trực tuyến nhờ các dữ liệu đo đạc, mà còn không cần sử dụng mô hình động học đầy đủ và chính xác của hệ thống. Điều này có ý nghĩa lớn trong thực tế khi việc thu được mô hình đủ chính xác của các hệ thống là rất khó khăn, chưa kể các thông số trong hệ thống có thể thay đổi trong quá trình hoạt động. Một số nghiên cứu khác mở rộng cho các hệ bị tác động bởi nhiễu ngoài, như các phương pháp quy hoạch động thích nghi bền vững [4] hoặc các phương pháp cho hệ có dạng multi-player zero-sum game [3], [11], hay các phương pháp khác xét đến ràng buộc đầu vào được trình bày trong [12], [13]. Một số nghiên cứu khác kết hợp điều khiển tối ưu thích nghi với các phương pháp phi tuyến bền vững như điều khiển trượt để tận dụng ưu điểm của từng phương pháp [14].

Xe hai bánh tự cân bằng là một hệ thống có bản chất là không ổn định, thiếu cơ cấu chấp hành và phi tuyến. Khi xe chuyển động trong môi trường phụ thuộc nhiều vào yếu tố bên ngoài như lực ma sát giữa bánh xe và mặt đường, tác động của gió, độ nghiêng của mặt đường và tải của xe có thể thay đổi. Do đó mô hình toán của xe hai bánh chứa nhiều yếu tố bất định và khó điều khiển. Đã có nhiều phương pháp điều khiển kinh điển như PID và hiện đại như backstepping, điều khiển thích nghi, điều khiển phi tuyến, điều khiển tối ưu đã được áp dụng cho xe hai bánh tự cân bằng, tuy nhiên các phương pháp này phần lớn dựa vào mô hình toán của xe. Hơn nữa, phương pháp ADP vẫn chưa được nghiên cứu và áp dụng cho lớp đối tượng này. Đây là phương pháp điều khiển có thể áp dụng cho đối tượng bất định mà không cần dùng mô hình toán. Tuy nhiên, để áp dụng được cho xe hai bánh tự cân bằng thì không những phải lựa chọn được hàm chi phí và cấu trúc mạng nơron phù hợp mà còn phải tìm được luật điều khiển ban đầu chấp nhận được. Đây là động lực để chúng tôi tiến hành nghiên cứu này.

Trong bài báo này, chúng tôi áp dụng thuật toán quy hoạch động thích nghi cho hệ phi tuyến bất định hoàn toàn, đã được trình bày trong cuốn sách "Robust Adaptive Dynamic Programming" [4] của Yu Jiang và Zhong-Ping Jiang cho đối tượng XHBTCB. Chất lượng điều khiển được kiểm chứng thông qua mô phỏng số trên phần mềm MATLAB. Bài báo được cấu trúc thành các phần như sau. Trong phần 2, mô hình động lực học của XHBTCB, đối tượng điều khiển trong bài báo, được trình bày. Trong phần 3, cơ sở lý thuyết và thuật toán tối ưu dựa trên quy hoạch động thích nghi được trình bày. Sau đó, tính hội tụ và ổn định được đề cập trong phần 4. Trong phần 5, kết quả mô phỏng cho thuật toán áp dụng trên đối tượng XHBTCB được trình bày để kiểm chứng tính đúng đắn của phương pháp. Cuối cùng, kết luận và định hướng phát triển nghiên cứu được đưa ra trong phần 6.

2. Mô hình động lực học của XHBTCB

Trong bài bào này, mô hình toán học của xe hai bánh tự cân bằng (XHBTCB) dựa trên tài liệu tham khảo [15] được sử dụng để kiểm nghiệm thuật toán điều khiển. Cấu trúc vật lý của XHBTCB được mô tả trong Hình 2, và định nghĩa của các ký hiệu được liệt kê trong Bảng 1.



Hình 1: Cấu trúc vật lý của XHBTCB

Bảng 1: Các ký hiệu, định nghĩa của XHBTCB

Ký hiệu	Định nghĩa
x	Vị trí xe hai bánh tự cân bằng
θ	Góc nghiêng của thân xe
ψ	Góc hướng của xe
d	Khoảng cách giữa bánh xe trái và bánh xe phải
l	Khoảng cách từ khối tâm thân xe đến trục nối hai bánh xe
r	Bán kính bánh xe
m_B	Khối lượng thân xe
m_W	Khối lượng bánh xe trái (phải)
J	Mômen quán tính của bánh xe ứng với trục bánh xe
K	Mômen quán tính của bánh xe ứng với trục thẳng đứng
K_m	Hằng số mômen xoắn
i_L, i_R	Dòng điện đi qua động cơ của bánh xe trái và động cơ của bánh xe phải
T_L, T_R	Mômen xoắn của động cơ của bánh xe trái và động cơ của bánh xe phải
γ_L, γ_R	Góc xoay của bánh xe trái và bánh xe phải
c_α	Hệ số ma sát nhớt trên trục bánh xe
I_1, I_2, I_3	Mômen quán tính của thân xe ứng với hệ quy chiếu {B}

Các phương trình chuyển động của hệ XHBTCB được cho như sau:

$$\left\{m_B + 2m_W + \frac{2J}{r^2}\right\}\ddot{x} - m_B l\left(\dot{\psi}^2 + \dot{\theta}^2\right)\sin\theta + \left(m_B l\cos\theta\right)\ddot{\theta} + \frac{2}{r}c_\alpha\left(\frac{\dot{x}}{r} - \dot{\theta}\right) = \frac{K_m\left(i_L + i_R\right)}{r} \quad (1)$$

$$\left(I_2 + m_B l^2\right)\ddot{\theta} + \left(m_B l\cos\theta\right)\ddot{x} + \left(I_3 - I_1 - m_B l^2\right)\dot{\psi}^2\sin\theta\cos\theta - m_B gl\sin\theta - 2c_\alpha\left(\frac{\dot{x}}{r} - \dot{\theta}\right) = -K_m\left(i_L + i_R\right) \quad (2)$$

$$\left\{I_3 + 2K + m_W\frac{d^2}{2} + J\frac{d^2}{2r^2}\right\}\ddot{\psi} + \left\{m_B l\dot{x} - 2\left(I_3 - I_1 - m_B l^2\right)\dot{\theta}\cos\theta\right\}\dot{\psi}\sin\theta + c_\alpha\dot{\psi}\frac{d^2}{2r^2} = \left(i_R - i_L\right)K_m\frac{d}{2r}. \quad (3)$$

Trong các phương trình động lực học hệ thống (1), (2), và (3), dòng điện phần ứng của các động cơ một chiều được coi là đầu vào của hệ thống, thay vì mômen như trong [15].

Ta định nghĩa các véctơ biến trạng thái và đầu vào như sau:

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{bmatrix}^T = \begin{bmatrix} x & \theta & \psi & \dot{x} & \dot{\theta} & \dot{\psi} \end{bmatrix}^T$$

$$\mathbf{u} = \begin{bmatrix} u_1 & u_2 \end{bmatrix}^T = \begin{bmatrix} i_L & i_R \end{bmatrix}^T$$

Khi đó, phương trình động lực học mô tả XHBTCB (1), (2), (3) có thể được viết lại dưới dạng ma trận như sau:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) + \mathbf{g}(\mathbf{x})\mathbf{u} = \mathbf{F}(\mathbf{x}, \mathbf{u}) \quad (4)$$

trong đó:

$$\dot{x}_1 = x_4, \dot{x}_2 = x_5, \dot{x}_3 = x_6, \dot{x}_4 = \mathbf{F}_4(\mathbf{x}, \mathbf{u}) = \frac{\Omega_1 + \Omega_2}{\Omega},$$

$$\dot{x}_5 = \mathbf{F}_5(\mathbf{x}, \mathbf{u}) = \frac{\Omega_3 + \Omega_4}{\Omega}, \dot{x}_6 = \mathbf{F}_6(\mathbf{x}, \mathbf{u}) = \frac{\Omega_5}{\Omega_6}$$

$$\Omega_1 = r^2\left(m_B l^2 + I_2\right)\left\{K_m\frac{u_1 + u_2}{r} + \frac{2c_\alpha}{r}\left(x_5 - \frac{x_4}{r}\right) + m_B l\sin\left(x_2\right)\left(x_5^2 + x_6^2\right)\right\}$$

$$\Omega_2 = m_B l r^2\cos\left(x_2\right)\left\{-\cos\left(x_2\right)\sin\left(x_2\right)\left(m_B l^2 + I_1 - I_3\right)x_6^2 + K_m\left(u_1 + u_2\right) + 2c_\alpha\left(x_5 - \frac{x_4}{r}\right) - m_B gl\sin\left(x_2\right)\right\}$$

$$\Omega_3 = \left[2J + \left(m_B + 2m_W\right)r^2\right]\left\{\cos\left(x_2\right)\sin\left(x_2\right)\left(m_B l^2 + I_1 - I_3\right)x_6^2 - K_m\left(u_1 + u_2\right) - 2c_\alpha\left(x_5 - \frac{x_4}{r}\right) + m_B gl\sin\left(x_2\right)\right\}$$

$$\Omega_4 = -m_B l r^2\cos\left(x_2\right)\left\{K_m\frac{u_1 + u_2}{r} + \frac{2c_\alpha}{r}\left(x_5 - \frac{x_4}{r}\right) + m_B l\sin\left(x_2\right)\left(x_5^2 + x_6^2\right)\right\}$$

$$\Omega_5 = -2r^2\left\{K_m d\frac{u_1 - u_2}{2r} + \frac{c_\alpha d^2 x_6}{2r^2} + x_6\sin\left(x_2\right)\left[m_B l x_4 + 2x_5\cos\left(x_2\right)\left(m_B l^2 + I_1 - I_3\right)\right]\right\}$$

$$\Omega_6 = \left\{2I_3 + 4K + m_W d^2 + 2\left(I_1 - I_3 + m_B l^2\right)\sin^2\left(x_2\right)\right\}r^2 + Jd^2$$

$$\Omega = \left(m_B l r\right)^2 \left[1 - \cos^2\left(x^2\right)\right] + 2I_2 J + 2Jm_B l^2 + \left(I_2 m_B + 2I_2 m_W + 2m_B m_W l^2\right) r^2.$$

3. Thuật toán điều khiển tối ưu dựa trên quy hoạch động thích nghi

Trong phần này, thuật toán quy hoạch động bán toàn cục cho hệ phi tuyến được phát triển và trình bày, dựa trên tài liệu tham khảo [4], [16].

3.1. Cơ sở lý thuyết

Xét hệ phi tuyến affine như sau:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u} \quad (5)$$

với $\mathbf{x} \in \mathbf{R}^n$ là véctơ trạng thái của hệ thống, $\mathbf{u} \in \mathbf{R}^m$ là véctơ tín hiệu điều khiển, $\mathbf{F}(\mathbf{x}) : \mathbf{R}^n \to \mathbf{R}^n$ và $\mathbf{G}(\mathbf{x}) : \mathbf{R}^n \to \mathbf{R}^{n\times m}$ là các ánh xạ liên tục Lipschitz trên một tập $\Omega \subset \mathbf{R}^n$ gồm gốc tọa độ, với $\mathbf{F}(\mathbf{0}) = \mathbf{0}$.

Ở đây, ta lưu ý rằng tính ổn định toàn cục tiệm cận được đảm bảo cho hệ tuyến tính, nhưng nói chung đối với hệ phi tuyến, tính chất này khó được đảm bảo [8]. Do đó, cơ sở lý thuyết của phương pháp chỉ được giới hạn trong trường hợp tính ổn định tiệm cận được thỏa mãn trong miền $\Omega \subset \mathbf{R}^n$.

Hàm chi phí ứng với một luật điều khiển $\mathbf{u}$ sẽ là:

$$V(\mathbf{x}) = \int_0^\infty r(\mathbf{x}, \mathbf{u})\, dt \quad (6)$$

với $r(\mathbf{x}, \mathbf{u}) = q(\mathbf{x}) + \mathbf{u}^T \mathbf{R}\mathbf{u}$ được lựa chọn là một hàm có dạng toàn phương để đảm bảo luật điều khiển tối ưu có thể xác định rõ ràng.

Trước khi giải bài toán điều khiển tối ưu, ta đặt ra giả thiết như sau:

Giả thiết 1: Giả thiết tồn tại một luật điều khiển phản hồi ổn định tiệm cận toàn cục $\mathbf{u}_0$ tại gốc tọa độ, trên một miền Ω cho hệ (5) với hàm chi phí (6) tương ứng là hữu hạn. Một luật điều khiển thỏa mãn giả thiết trên được gọi là luật điều khiển ổn định chấp nhận được [17].

Bài toán điều khiển tối ưu bây giờ có thể được phát biểu như sau: Xét hệ phi tuyến liên tục (5) và tập hợp các luật điều khiển chấp nhận được $\Psi(\Omega)$, tìm luật điều khiển để tối thiểu hóa hàm chi phí (6).

Ta định nghĩa C^1 là tập hợp các hàm liên tục khả vi và P^1 là tập tất cả các hàm trong C^1 xác định dương và thỏa mãn $\|x\| \to \infty$ thì $f(x) \to \infty$. Khi đó ta nhận thấy hàm $V(\mathbf{x})$ trong công thức (6) phải thuộc tập C^1, nói cách khác:

$$\left(\nabla V(\mathbf{x})\right)^T \left(\mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}\right) + r(\mathbf{x}, \mathbf{u}) = 0, V(\mathbf{0}) = 0 \quad (7)$$

Phương trình (7) còn được gọi là phương trình Lyapunov cho hệ phi tuyến. Định nghĩa hàm Hamilton như sau:

$$H\left(\mathbf{x}, \mathbf{u}, V\right) = r(\mathbf{x}, \mathbf{u}) + \left(\nabla V(\mathbf{x})\right)^T \left(\mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}\right) \quad (8)$$

và hàm chi phí tối ưu $V^*(\mathbf{x})$ thỏa mãn phương trình HJB:

$$\min_{u \in \Psi(\Omega)} H\left(\mathbf{x}, \mathbf{u}, V^*\right) = 0 \quad (9)$$

Giả thiết rằng tồn tại duy nhất $V^* \in P^1$ là nghiệm của phương trình HJB (9), thì luật điều khiển tối ưu được xác định bởi công thức:

$$\mathbf{u}^*(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{G}^T(\mathbf{x})\nabla V^*(\mathbf{x}) \quad (10)$$

có thể ổn định tiệm cận toàn cục hệ (5) tại $\mathbf{x} = 0$.

Nếu xác định được một hàm thuộc lớp P^1 là nghiệm của phương trình HJB (9) thì ta có thể tìm được công thức tường minh của luật điều khiển tối ưu. Tuy nhiên, phương trình HJB phi tuyến nói chung là rất khó để giải. Do đó, cũng giống như với hệ tuyến tính, phương pháp lặp cũng đã được phát triển cho hệ phi tuyến, cụ thể như sau.

Định lý 1: Cho $\mathbf{u}_0$ là luật điều khiển ổn định tiệm cận toàn cục tại gốc tọa độ của hệ (5) (Giả thiết 1). Khi đó, với $k = 0, 1, \ldots$, hàm chi phí $V_k(\mathbf{x}) \in C^1$ thu được bằng việc giải phương trình:

$$\nabla V_k^T(\mathbf{x})\left[\mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}_k\right] + r(\mathbf{x}, \mathbf{u}_k) = 0 \quad (11)$$

và luật điều khiển $\mathbf{u}_k$ được tính toán đệ quy theo công thức:

$$\mathbf{u}_{k+1}(\mathbf{x}) = -\frac{1}{2}\mathbf{R}^{-1}\mathbf{G}^T(\mathbf{x})\nabla V_k(\mathbf{x}) \quad (12)$$

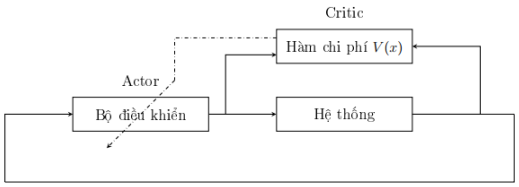
Khi đó, ta có các tính chất sau:

- $V^*(\mathbf{x}) \le V_{k+1}(\mathbf{x}) \le V_k(\mathbf{x}), \forall \mathbf{x} \in \mathbf{R}^n$
- $\mathbf{u}_k$ là luật điều khiển ổn định toàn cục.
- Đặt $\lim_{k\to\infty} V_k(\mathbf{x}_0) = V(\mathbf{x}_0)$ và $\lim_{k\to\infty} \mathbf{u}_k(\mathbf{x}_0) = \mathbf{u}(\mathbf{x}_0)$ với $\forall \mathbf{x}_0 \in \mathbf{R}^n$. Khi đó, $V^* = V$ và $\mathbf{u}^* = \mathbf{u}$ nếu $V \in C^1$.

Chứng minh: Xem tài liệu tham khảo [4].

3.2. Thuật toán

Trong phần này, phương pháp lặp PI để xấp xỉ nghiệm của phương trình HJB và luật điều khiển tối ưu trên cơ sở mạng nơron, đã được đề xuất trong [4], được trình bày. Phương pháp là phiên bản mở rộng của phương pháp cho hệ tuyến tính được trình bày trong [7]. Thuật toán lặp PI, cũng giống các thuật toán học tăng cường khác, có thể được thực thi trực tuyến trên cấu trúc Actor-Critic [9]. Cấu trúc trên được minh họa trong Hình 2. Trong cấu trúc Actor-Critic, dựa trên khả năng xấp xỉ bất kỳ hàm phi tuyến trơn trên một tập compact của mạng nơron, hàm chi phí $V_k(\mathbf{x})$ và luật điều khiển $\mathbf{u}_{k+1}(\mathbf{x})$ được xấp xỉ bằng hai mạng nơron, được gọi tương ứng là mạng nơron Critic và mạng nơron Actor.



Hình 2: Cấu trúc Actor-Critic

Với mỗi $k = 0,1,\ldots$, hàm V_k và luật điều khiển $\mathbf{u}_k$ được xấp xỉ trên miền Ω như sau:

$$\begin{aligned}\hat{V}_k(\mathbf{x}) &= \mathbf{c}_k^T\phi(\mathbf{x})\\ \hat{\mathbf{u}}_k(\mathbf{x}) &= \mathbf{w}_k^T\psi(\mathbf{x})\end{aligned} \tag{13}$$

trong đó $\phi(\mathbf{x}):\mathbf{R}^n \to \mathbf{R}^{N_1}$ và $\psi(\mathbf{x}):\mathbf{R}^n \to \mathbf{R}^{N_2}$, với N_1 và N_2 là các số nguyên dương đủ lớn, là véctơ các hàm trơn độc lập tuyến tính trên miền Ω và bằng 0 tại $\mathbf{x}=0$, $\mathbf{c}_k \in R^{N_1}$ và $\mathbf{w}_k \in R^{N_2\times m}$ là véctơ hoặc ma trận trọng số được cập nhật. Nói cách khác, với mạng Critic, ta sử dụng một mạng nơron với N_1 nơron ở lớp ẩn và hàm kích hoạt $\phi(\mathbf{x})$, trọng số của lớp ẩn được coi đều bằng 1 và không thay đổi trong suốt quá trình huấn luyện. Đầu ra của mạng có hàm kích hoạt là hàm tuyến tính, với véctơ trọng số là $\mathbf{c}_k$. Tương tự với mạng nơron Actor m đầu ra dùng để xấp xỉ $\mathbf{u}_k$.

Ta viết lại phương trình (5) dưới dạng như sau:

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u_k} + \mathbf{G}(\mathbf{x})(\mathbf{u}-\mathbf{u_k}) \tag{14}$$

Xét đạo hàm của $V_k(\mathbf{x})$, kết hợp với (6) và (12) ta có:

$$\begin{aligned}\dot{V}_k &= \nabla V_k(\mathbf{x})\left[\mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}_k + \mathbf{G}(\mathbf{x})(\mathbf{u}-\mathbf{u}_k)\right]\\ &= -q(\mathbf{x}) - \mathbf{u}_k^T\mathbf{R}\mathbf{u}_k + \nabla V_k(\mathbf{x})\mathbf{G}(\mathbf{x})(\mathbf{u}-\mathbf{u}_k)\\ &= -q(\mathbf{x}) + \mathbf{u}_k^T\mathbf{R}\mathbf{u}_k - 2\mathbf{u}_{k+1}^T\mathbf{R}(\mathbf{u}-\mathbf{u}_k)\end{aligned} \tag{15}$$

Lấy tích phân trong công thức (15) trong khoảng thời gian $[t, t+T]$, ta có:

$$\begin{aligned}&V_k\left(\mathbf{x}(t+T)\right) - V_k\left(\mathbf{x}(t)\right)\\ &= -\int_t^{t+T} q(\mathbf{x}) + \mathbf{u}_k^T\mathbf{R}\mathbf{u}_k + 2(\mathbf{u}-\mathbf{u}_k)^T\mathbf{R}\mathbf{u}_{k+1}\,d\tau\end{aligned} \tag{16}$$

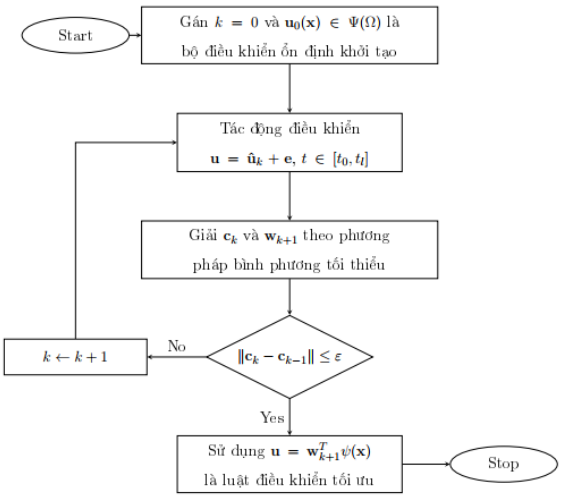
trong đó $\mathbf{u} = \mathbf{u}_k + \mathbf{e}$ là tín hiệu đầu vào tác động lên hệ thống trong khoảng thời gian $[t, t+T]$, với $\mathbf{e}$ là tín hiệu nhiễu thăm dò biên độ nhỏ.

Thay thế V_k, $\mathbf{u}_k$ và $\mathbf{u}_{k+1}$ trong (16) bằng xấp xỉ mạng nơron trong (13) ta có:

$$\begin{aligned}&\mathbf{c}_k^T\left[\phi\left(\mathbf{x}(t)\right) - \phi\left(\mathbf{x}(t+T)\right)\right]\\ &-2\int_t^{t+T}(\mathbf{u}-\hat{\mathbf{u}}_k)^T\mathbf{R}\mathbf{w}_{k+1}^T\psi(\mathbf{x})\,d\tau\\ &= \int_t^{t+T} q(\mathbf{x}) + \hat{\mathbf{u}}_k^T\mathbf{R}\hat{\mathbf{u}}_k\,d\tau + e_k\end{aligned} \tag{17}$$

với e_k là tổng sai lệch gây ra bởi xấp xỉ mạng nơron.

Áp dụng thuật toán lặp PI, ta có thể giải được các trọng số mạng nơron $\hat{\mathbf{c}}_k$ và $\hat{\mathbf{w}}_{k+1}$ bằng phương pháp tối thiểu hóa hàm sai lệch $\sum_{i=1}^{l} e_{k,l}^{\ 2}$ trong (17) sau khi thu thập đủ dữ liệu của của các khoảng thời gian lấy mẫu $[t_0,t_1],[t_1,t_2]\ldots,[t_{l-1},t_l]$. Lưu đồ thuật toán lặp PI cho hệ phi tuyến được mô tả trong Hình 3.



Hình 3: Lưu đồ thuật toán lặp PI cho hệ phi tuyến

4. Tính ổn định và hội tụ

Trong phần này, tính hội tụ của thuật toán và tính ổn định của hệ kín sẽ được xem xét. Đầu tiên, ta đặt ra các giả thiết sau.

Giả thiết 2: Giả thiết tồn tại số tự nhiên l_0 và $\delta > 0$ sao cho với mọi $l \ge l_0$ ta có:

$$\frac{1}{l}\sum_{i=0}^{l}\theta_{k,i}^T\theta_{k,i} \ge \delta I_{N_1+N_2}$$

với

$$\theta_{k,i}^T = \begin{bmatrix}\phi_1\left(\mathbf{x}(t+T)\right) - \phi_1\left(\mathbf{x}(t)\right)\\ \vdots\\ \phi_{N_1}\left(\mathbf{x}(t+T)\right) - \phi_{N_1}\left(\mathbf{x}(t)\right)\\ \int_t^{t+T}(\mathbf{u}-\mathbf{u}_k)^T\mathbf{R}\mathbf{w}_{k+1}^T\psi_1(\mathbf{x})\,d\tau\\ \vdots\\ \int_t^{t+T}(\mathbf{u}-\mathbf{u}_k)^T\mathbf{R}\mathbf{w}_{k+1}^T\psi_{N_2}(\mathbf{x})\,d\tau\end{bmatrix} \in \mathbf{R}^{N_1+N_2}$$

Giả thiết 3: Giả thiết hệ kín (5) là ổn định ISS khi nhiễu thăm dò được áp dụng vào luật điều khiển.

Định lý 2: Với các giả thiết 2 và 3, với mọi $k \ge 0$ và giá trị $\epsilon > 0$ cho trước, tồn tại các số nguyên dương k^*, N_1^* và N_2^* thỏa mãn:

$$\begin{aligned}\left|\mathbf{c}_k^T\phi(\mathbf{x}) - V^*(\mathbf{x})\right| &< \varepsilon\\ \left|\mathbf{w}_k^T\psi(\mathbf{x}) - \mathbf{u}^*(\mathbf{x})\right| &< \varepsilon\end{aligned} \tag{18}$$

với mọi $\mathbf{x} \in \Omega$, $N_1 > N_1^*$ và $N_2 > N_2^*$.

Chứng minh: Xem tài liệu tham khảo [4].

Một cách nói chung, mạng nơron không có khả năng xấp xỉ các hàm phi tuyến trên toàn không gian trạng thái $\mathbf{R}^n$ mà chỉ trên một tập compact. Do đó, mặc dù thuật toán được nghiên cứu đã đảm bảo được tính hội tụ nhưng luật điều khiển thu được vẫn có thể không áp dụng được nếu trạng thái của hệ thống vượt ra ngoài tập compact Ω, từ đó gây ra mất ổn định. Do đó, trong [4] các tác giả đã đưa ra định lý sau để phân tích tính ổn định của hệ kín.

Định lý 3: Với các giả thiết 1, 2 và 3, hệ kín sẽ ổn định

tiệm cận tại gốc tọa độ nếu:

$$q(\mathbf{x}) > \left(\mathbf{u}_{k+1} - \hat{\mathbf{u}}_{k+1}\right)^T \mathbf{R}\left(\mathbf{u}_{k+1} - \hat{\mathbf{u}}_{k+1}\right), \forall \mathbf{x} \in \Omega \setminus \{0\} \quad (19)$$

Chứng minh: Với luật điều khiển $\mathbf{u} = \hat{\mathbf{u}}_{k+1}$, ta có đạo hàm của hàm Lyapunov V_k trở thành:

$$\begin{aligned}
\dot{V}_k &= \nabla V_k^T(\mathbf{x})\left(\mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\hat{\mathbf{u}}_{k+1}\right) \\
&= \nabla V_k^T(\mathbf{x})\left(\mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x})\mathbf{u}_{k+1}\right) \\
&\quad + \nabla V_k^T(\mathbf{x})\mathbf{G}(\mathbf{x})\left(\hat{\mathbf{u}}_{k+1} - \mathbf{u}_{k+1}\right) \\
&= -q(\mathbf{x}) - \mathbf{u}_{k+1}^T \mathbf{R}\mathbf{u}_{k+1} \\
&\quad - \left(\mathbf{u}_{k+1} - \mathbf{u}_k\right)^T \mathbf{R}\left(\mathbf{u}_{k+1} - \mathbf{u}_k\right) \\
&\quad - 2\mathbf{u}_{k+1}^T \mathbf{R}\left(\hat{\mathbf{u}}_{k+1} - \mathbf{u}_{k+1}\right) \\
&\le -q(\mathbf{x}) + \mathbf{u}_{k+1}^T \mathbf{R}\mathbf{u}_{k+1} - 2\mathbf{u}_{k+1}^T \mathbf{R}\left(\hat{\mathbf{u}}_{k+1} - \mathbf{u}_{k+1}\right) \\
&\le -q(\mathbf{x}) + \left(\hat{\mathbf{u}}_{k+1} - \mathbf{u}_{k+1}\right)^T \mathbf{R}\left(\hat{\mathbf{u}}_{k+1} - \mathbf{u}_{k+1}\right)
\end{aligned} \quad (20)$$

với $\forall \mathbf{x} \in \Omega \setminus \{0\}$. Nên nếu điều kiện (19) được thỏa mãn thì hệ kín sẽ ổn định tiệm cận tại gốc tọa độ.

Do đó, thuật toán được trình bày cho hệ phi tuyến được các tác giả gọi là quy hoạch động thích nghi bán toàn cục [4].

Lưu ý: Lựa chọn cấu trúc mạng cho mạng nơron dùng để xấp xỉ hàm V_k và luật điều khiển $\mathbf{u}_k$ vẫn là một vấn đề mở chưa được đề cập trong các công trình nghiên cứu trước đây. Trong bài báo này, các hàm kích hoạt $\phi_j(\mathbf{x})$ được chọn có dạng toàn phương, trong khi đó $\psi_j(\mathbf{x})$ được lựa chọn từ các phần tử độc lập tuyến tính của bộ điều khiển ban đầu ổn định hệ thống $\mathbf{u}_0$.

5. Mô phỏng kiểm chứng

Trong phần này, thuật toán tối ưu dựa trên quy hoạch động cho hệ phi tuyến đã trình bày được áp dụng cho hệ XHBTCB và kiểm chứng thông qua mô phỏng số trên phần mềm MATLAB. Các thông số của đối tượng thu được từ mô hình trong phòng thí nghiệm như sau: $m_B = 0.5(kg)$, $m_W = 0.04(kg)$, $l = 0.08(m)$, $d = 0.16(m)$, $r = 0.033(m)$, $g = 9.81(m/s^2)$, $c_\alpha = 5.10^{-4}(Ns/m)$, $K_m = 0.412\ (Nm/A)$.

Hàm chi phí trong bài toán điều khiển tối ưu được định nghĩa như sau:

$$J(\mathbf{x}, \mathbf{u}) = \int_0^\infty \mathbf{x}^T \mathbf{Q}\mathbf{x} + \mathbf{u}^T \mathbf{R}\mathbf{u}\, d\tau \quad (21)$$

với $\mathbf{Q} = \mathrm{diag}(1, 0.5, 2, 0.05, 0.05, 1)$ và $\mathbf{R} = \mathbf{I}_2$. Thời gian lấy mẫu là 0.01 s, thuật toán được thực thi sau 200 mẫu dữ liệu, tương đương với sau mỗi 2 s. Tín hiệu nhiễu thăm dò được lựa chọn là dạng tổng các tín hiệu sin như sau [4]:

$$e = 0.1\sum \sin(\omega_i t)$$

trong đó ω_i với $i = 1, ..., 100$ là tần số được chọn ngẫu nhiên trong khoảng $[-500, 500]$.

Mạng nơron được sử dụng có cấu trúc như sau: $N_1 = 21$, $\phi(\mathbf{x}) = \left[x_i x_j\right]_{i,j=1,..6}^T$, $N_2 = 6$, $\psi(\mathbf{x}) = \left[\dfrac{x_i \cos(x_2)}{1 + \sin^2(x_2)}\right]_{i=1,..6}^T$ và thông số khởi tạo mạng

$$\mathbf{w}_0 = \begin{bmatrix} 0.2 & 0.6 & 0.3 & 0.2 & 0.1 & 0.2 \\ 0.2 & 0.6 & -0.3 & 0.2 & 0.1 & -0.2 \end{bmatrix}^T$$

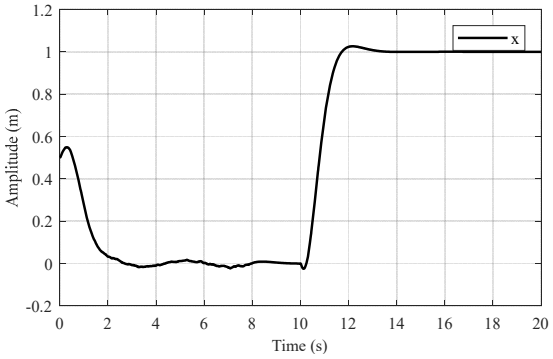
Trong mô phỏng này, ta xét chuyển động trên đường thẳng của xe, nhiệm vụ điều khiển là đảm bảo cho xe bám vị trí đặt, trong khi góc nghiêng thân xe và góc hướng được giữ càng nhỏ càng tốt và tiến về 0 ở trạng thái xác lập. Cụ thể, ta giả sử xe chuyển động từ vị trí ban đầu 0.5 (m) về gốc tọa độ trong 10 giây đầu tiên, rồi di chuyển tới vị trí đặt mới 1 (m) trong 10 giây tiếp theo.

Bộ điều khiển tối ưu được tìm ra đảm bảo cho hệ bám với giá trị đặt. Các véctơ trọng số tối ưu của mạng nơron Critic và Actor thu được từ thuật toán sau 4 vòng lặp như sau:

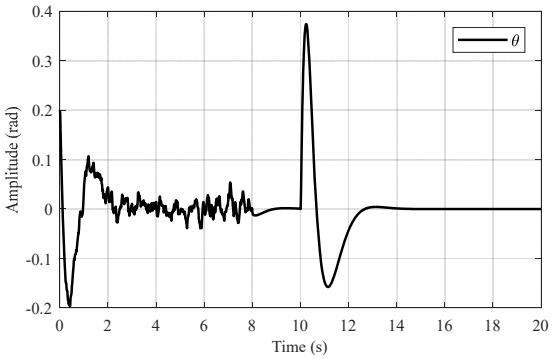
$$\mathbf{c}_3 = \begin{bmatrix} 0.075 & 0.069 & -0.030 & ... & 0 & 0.002 \end{bmatrix}^T$$

$$\mathbf{w}_4 = \begin{bmatrix} 0.19 & 0.62 & 0.39 & 0.18 & 0.08 & 0.23 \\ 0.26 & 0.67 & -0.55 & 0.22 & 0.08 & -0.23 \end{bmatrix}^T$$

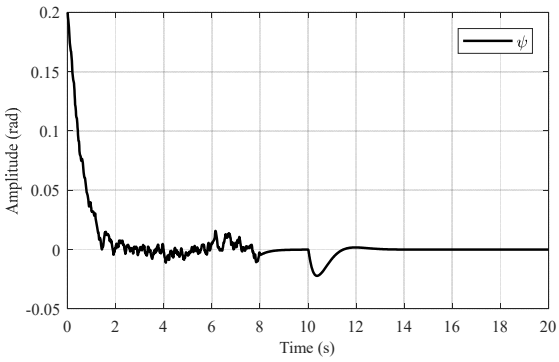
Kết quả mô phỏng với thuật toán tối ưu phi tuyến dựa trên quy hoạch động thích nghi được thể hiện trong các Hình 4, 5 và 6, lần lượt cho dịch chuyển của xe, góc nghiêng θ của thân xe và góc hướng ψ của xe. Như có thể thấy, thuật toán tìm ra bộ điều khiển tối ưu sau quá trình học 8 giây, và bộ điều khiển tối ưu thu được đảm bảo cho hệ ổn định.



Hình 4: Dịch chuyển của xe



Hình 5: Góc lắc thân xe



Hình 6: Góc hướng của xe

6. Kết luận

Bài báo đã tìm hiểu thuật toán điều khiển tối ưu dựa trên quy hoạch động thích nghi [4]. Thuật toán điều khiển ứng dụng quy hoạch động thích nghi cho hệ phi tuyến mô hình bất định hoàn toàn và không phụ thuộc thời gian được trình bày chi tiết. Sau đó, thuật toán đã được áp dụng cho mô hình XHBTCB và kiểm chứng chất lượng bộ điều khiển thông qua mô phỏng số trên phần mềm MATLAB. Thuật toán quy hoạch đông thích nghi được áp dụng đã giải quyết tốt yêu cầu đặt ra đó là tìm lời giải trực tuyến cho bài toán điều khiển tối ưu các hệ thống động học khi mô hình toán học của hệ thống được coi là bất định. Tuy nhiên, vấn đề còn tồn tại đó là thuật toán chỉ là ổn định bán toàn cục, theo nghĩa hệ kín sẽ ổn định nếu một số điều kiện nhất định được thỏa mãn. Hơn nữa, việc lựa chọn cấu trúc mạng nơron và bộ trọng số mạng khởi tạo để đảm bảo hệ không mất ổn định trong quá trình học cũng chưa được phân tích chặt chẽ. Đó cũng chính là dự định phát triển về mặt lý thuyết trong tương lai. Cuối cùng, định hướng phát triển về thực nghiệm là áp dụng các phương pháp này trên mô hình xe thực trong phòng thí nghiệm.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi Trường Đại học Bách khoa Hà Nội trong đề tài mã số T2018-PC-052.

Tài liệu tham khảo

[1] D. Vrabie, “Online adaptive optimal control for continuous-time systems,” 2010.

[2] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*. 1998.

[3] K. G. Vamvoudakis, “Online learning algorithms for differential dynamic games and optimal control,” 2011.

[4] Y. Jiang and Z.-P. Jiang, *Robust adaptive dynamic programming*. 2017.

[5] J. J. Murray, C. J. Cox, G. G. Lendaris, and R. Saeks, *Adaptive dynamic programming*. 2002.

[6] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, “Adaptive optimal control for continuous-time linear systems based on policy iteration,” *Automatica*, vol. 45, no. 2, pp. 477–484, 2009.

[7] Y. Jiang and Z.-P. Jiang, “Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics,” *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[8] D. Vrabie and F. Lewis, “Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems,” *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

[9] K. G. Vamvoudakis and F. L. Lewis, “Online actor--critic algorithm to solve the continuous-time infinite horizon optimal control problem,” *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[10] Z.-P. Jiang, Yu and Jiang, “Global adaptive dynamic programming for continuous-time nonlinear systems,” *IEEE Trans. Automat. Contr.*, vol. 60, no. 11, pp. 2917–2929, 2015.

[11] D. Liu, Q. Wei, D. Wang, X. Yang, and H. Li, *Adaptive dynamic programming with applications in optimal control*. 2017.

[12] T. Cheng, F. L. Lewis, and M. Abu-Khalaf, “Fixed-final-time-constrained optimal control of nonlinear systems using neural network HJB approach,” *IEEE Trans. Neural Networks*, vol. 18, no. 6, pp. 1725–1737, 2007.

[13] D. Liu, D. Wang, and X. Yang, “An iterative adaptive dynamic programming algorithm for optimal control of unknown discrete-time nonlinear systems with constrained inputs,” *Inf. Sci. (Ny).*, vol. 220, pp. 331–342, 2013.

[14] Q.-Y. Fan and G.-H. Yang, “Adaptive actor--critic design-based integral sliding-mode control for partially unknown nonlinear systems with input disturbances,” *IEEE Trans. neural networks Learn. Syst.*, vol. 27, no. 1, pp. 165–177, 2015.

[15] S. Kim and S. Kwon, “Dynamic modeling of a two-wheeled inverted pendulum balancing mobile robot,” *Int. J. Control. Autom. Syst.*, vol. 13, no. 4, pp. 926–933, 2015.

[16] Y. Jiang and Z.-P. Jiang, “Robust approximate dynamic programming and global stabilization with nonlinear dynamic uncertainties,” in *50th IEEE Conference on Decision and Control and European Control Conference*, 2011, pp. 115–120.

[17] R. W. Beard, G. N. Saridis, and J. T. Wen, “Galerkin approximations of the generalized Hamilton-Jacobi-Bellman equation,” *Automatica*, vol. 33, no. 12, pp. 2159–2177, 1997.