Q-Learning for Inventory Management: an application case

# Abstract

Các ứng dụng gần đây của công nghệ 4.0 đang dần dẫn đến việc giới thiệu và phát triển trí tuệ nhân tạo trong các hệ thống sản xuất. Nghiên cứu này thảo luận về việc đánh giá một mô hình Học máy (Machine Learning), dựa trên Học tăng cường (Reinforcement Learning), có thể cho phép tối ưu hóa mức tồn kho ở cấp độ máy móc, từ đó cải thiện hệ thống đặt hàng và quản lý hàng tồn kho. Mô hình này được áp dụng cho một trường hợp công nghiệp thực tế và kết quả, khi so sánh với những kết quả thu được bằng cách sử dụng các kỹ thuật tối ưu hóa truyền thống, cho thấy một hiệu suất đáng kể. Có thể kết luận rằng mô hình Học máy được phát triển có thể được sử dụng thành công để cải thiện chu kỳ đặt hàng của một doanh nghiệp.

# 1. Introduction

Trong những năm gần đây, các công ty đã phát triển sự quan tâm ngày càng tăng đối với các vấn đề liên quan đến quản lý hàng tồn kho. Để tồn tại trước những thách thức của các thị trường mới, ngày càng năng động và cạnh tranh, các công ty đã chú trọng đến một số khía cạnh của quản lý kinh doanh mà trước đây thường bị bỏ qua, đặc biệt bằng cách sử dụng các công nghệ mới do Công nghiệp 4.0 cung cấp. Trên thực tế, các nghiên cứu mới đã chứng minh rằng việc thiết kế bố cục chính xác và quản lý hàng tồn kho hiệu quả có tác động mạnh mẽ đến hiệu suất của công ty, chẳng hạn như mức độ dịch vụ, giảm chi phí hậu cần và nhìn chung là tăng khả năng cạnh tranh. Đại dịch coronavirus (COVID-19) đã làm tăng hiệu ứng bullwhip—tức là sự gia tăng biến động đơn hàng khi đơn hàng di chuyển ngược dòng trong chuỗi cung ứng—và dẫn đến những ràng buộc về con người, công nghệ và quản lý đối với các công ty sản xuất do khan hiếm tài nguyên hoặc gián đoạn chuỗi cung ứng [1]. Việc áp dụng các công nghệ Công nghiệp 4.0 [2] có thể giải quyết những thách thức do đại dịch gây ra (ví dụ: thiếu hụt nguồn cung nguyên liệu thô, thay đổi nhu cầu sản phẩm và đóng cửa nhà máy do phong tỏa chung) và ngăn chặn sự tái diễn của chúng trong trường hợp các sự kiện gián đoạn khác trong tương lai bằng cách cải thiện khả năng phục hồi của công ty và giảm thiểu rủi ro chuỗi cung ứng thông qua tính linh hoạt, độ tin cậy, sự mạnh mẽ và khả năng phản ứng ([3]; [4]; [5]).

Cuộc cách mạng Công nghiệp 4.0 đã bùng nổ trong những năm gần đây, tuy nhiên, nó thường gặp khó khăn trong việc tìm kiếm sự triển khai, đặc biệt là ở các thực tế địa phương **[16]**. Điều này chắc chắn là do cuộc cách mạng này đòi hỏi những khoản đầu tư khổng lồ, nhưng cũng bởi vì những lợi thế từ việc áp dụng các công nghệ này thường không thực sự được hiểu rõ, đặc biệt là đối với việc sử dụng chúng trong logistics và quản lý vật liệu. Tuy nhiên, sự phát triển của các kỹ thuật sản xuất phải đi kèm với việc cải tiến các kỹ thuật quản lý dòng vật liệu, nếu không sẽ khó có thể tận hưởng trọn vẹn những lợi thế có được từ việc đầu tư vào máy móc ngày càng tiên tiến. Cụ thể, các chính sách tồn kho truyền thống thể hiện hiệu suất kém hơn trong một bối cảnh đặc trưng bởi sự biến động và không chắc chắn lớn, chẳng hạn như môi trường sau đại dịch. Trên thực tế, trong bối cảnh như vậy, các thông số được thiết lập bên trong hệ thống thông tin nên được sửa đổi và cập nhật thủ công thường xuyên. **Học máy (Machine Learning)** có thể là một giải pháp cho vấn đề này: cụ thể là mô hình được đề xuất sẽ gợi ý một phương pháp tiếp cận mới để thiết lập các thông số của hệ thống thông tin liên quan đến quản lý hàng tồn kho.

**Học máy (Machine learning)** có nhiều ứng dụng, chẳng hạn như trong lý thuyết trò chơi, lý thuyết điều khiển, lý thuyết thông tin, nghiên cứu vận hành, tối ưu hóa dựa trên mô phỏng và các hệ thống đa tác tử. **Học tăng cường (Reinforcement learning)** là sự kết hợp của phương pháp Monte Carlo và lập trình động, sử dụng các quy trình giải pháp khác nhau, chẳng hạn như Q-learning hoặc thuật toán Sarsa , được biết đến là các giải pháp gần tối ưu, cho phép tìm ra các hành động cần thực hiện trong một trạng thái cụ thể để tối ưu hóa phần thưởng kỳ vọng. Tuy nhiên, có rất ít nghiên cứu toàn diện về việc tích hợp Học tăng cường (RL) vào kiểm soát và quản lý hàng tồn kho: nó chủ yếu được sử dụng để giải quyết các vấn đề khác nhau, ví dụ như tối ưu hóa các khoản đầu tư tài chính hoặc giảm chi phí vận chuyển, tăng khả năng hiển thị và phát triển sự linh hoạt trong quản lý vận chuyển hợp tác. Gần đây, nghiên cứu về ứng dụng của RL cũng đang lan rộng trong lĩnh vực lập kế hoạch và kiểm soát sản xuất của chuỗi cung ứng. RL đã được áp dụng chủ yếu cho các vấn đề lập lịch sản xuất, tiếp theo là quản lý mua hàng và cung ứng, và các thuật toán RL được sửa đổi nhiều nhất là mô hình độc lập (model-free) và tác nhân đơn lẻ (single-agent) được áp dụng cho các môi trường lập kế hoạch và kiểm soát sản xuất được đơn giản hóa. Chuyển từ góc độ công ty sang góc độ chuỗi cung ứng, quản lý hàng tồn kho là ứng dụng phổ biến nhất của Học tăng cường, vì nó là một yếu tố then chốt để đồng bộ hóa chuỗi cung ứng. Các mô hình giải quyết vấn đề quản lý hàng tồn kho thường sử dụng tác nhân RL để điều phối dòng vật liệu giữa nhiều địa điểm trong chuỗi cung ứng.

Những nghiên cứu này chủ yếu tập trung vào ba giai đoạn: **cung ứng, sản xuất và phân phối**. Trong một số trường hợp, sự chú ý của các nhà nghiên cứu đã được hướng tới việc phối hợp và tích hợp các chính sách quản lý tồn kho giữa nhiều hơn ba giai đoạn. Khi thiếu sự phối hợp như vậy giữa các tác nhân khác nhau, mỗi thực thể sẽ đưa ra quyết định dựa trên tiêu chí riêng của mình, dẫn đến việc **tối ưu hóa cục bộ** so với **tối ưu hóa toàn cầu**. Tuyệt vời. Đây là bản dịch tiếng Việt của đoạn văn bạn yêu cầu:

Những nghiên cứu này chủ yếu tập trung vào ba giai đoạn: **cung ứng, sản xuất và phân phối**. Trong một số trường hợp, sự chú ý của các nhà nghiên cứu đã được hướng tới việc phối hợp và tích hợp các chính sách quản lý tồn kho giữa nhiều hơn ba giai đoạn. Khi thiếu sự phối hợp như vậy giữa các tác nhân khác nhau, mỗi thực thể sẽ đưa ra quyết định dựa trên tiêu chí riêng của mình, dẫn đến việc **tối ưu hóa cục bộ** so với **tối ưu hóa toàn cầu**.

Thật vậy, ví dụ, [11] trình bày một phương pháp để quản lý các quyết định tồn kho ở tất cả các giai đoạn của chuỗi cung ứng một cách **tích hợp** thông qua quy trình quyết định Markov và sử dụng thuật toán học tăng cường để xác định một chính sách tồn kho gần tối ưu theo tiêu chí phần thưởng trung bình. Ngược lại, [12] trình bày một hệ thống kiểm soát hàng tồn kho, trong trường hợp thời gian giao hàng và nhu cầu không chắc chắn, để xác định các chính sách đặt hàng của từng lĩnh vực trong chuỗi cung ứng. Mô hình được đề xuất đã được xây dựng như một phương pháp học tăng cường nhằm **giảm thiểu tổng chi phí tồn kho** bao gồm chi phí quản lý và chi phí đơn hàng tồn đọng (backorder costs). [13] đã sử dụng học máy để tham số hóa Mô hình Vận hành theo Nhu cầu (Demand-Driven Operation Model) đối mặt với nhu cầu không xác định, và đặc biệt để điều chỉnh linh hoạt **ngưỡng tăng đột biến đơn hàng** và **chân trời tăng đột biến đơn hàng**. Một thuật toán học tăng cường với ba hàm phần thưởng khác nhau đã được ghép nối với mô hình mô phỏng dòng chảy Sản xuất theo Yêu cầu Nguyên vật liệu theo Nhu cầu (DDMRP) đối mặt với nhu cầu bất thường bao gồm các đợt tăng đột biến. [14] đã xác định một khoảng trống nghiên cứu là cần có nhiều phương pháp RL hơn nên được áp dụng và thử nghiệm trong lập kế hoạch tài nguyên cơ sở, lập kế hoạch năng lực, và **các vấn đề quản lý hàng tồn kho**, và nói chung là trong các **vấn đề thực tế** hơn.

Mục đích của nghiên cứu này do đó là phát triển một mô hình **Học máy (Machine Learning)** để **tối ưu hóa việc quản lý hàng tồn kho** đối mặt với một bối cảnh biến đổi và không chắc chắn. Nghiên cứu này đề cập đến vấn đề quản lý **mức tồn kho có sẵn trên máy móc**, xem xét lượng tiêu thụ nguyên liệu thô của máy và thời gian giao hàng, với mục tiêu đạt được một **chính sách đặt hàng tối ưu** cho các sản phẩm, sử dụng thuật toán **Q-learning**.

Sau phần giới thiệu này, **Mục 2** giới thiệu về khái niệm **Học tăng cường (Reinforcement Learning)**. **Mục 3** định nghĩa vấn đề và mô tả mô hình đã được phát triển. **Mục 4** trình bày các kết quả và thảo luận về mô phỏng trường hợp nghiên cứu được đề xuất. Trong mục cuối cùng, chúng tôi phân tích các kết quả thu được thông qua nghiên cứu, và các mở rộng tiềm năng trong tương lai sẽ được thảo luận.

# 2. Reinforcement Learning

Trí tuệ nhân tạo (Artificial intelligence) quan tâm đến việc tạo ra một chương trình máy tính thể hiện trí thông minh giống con người. Học máy (Machine learning) là một lĩnh vực con của trí tuệ nhân tạo, mang lại cho máy tính khả năng tự động học hỏi từ dữ liệu. **Học tăng cường (Reinforcement Learning - RL)** là nhiệm vụ học tập thông qua thử và lỗi, và mục tiêu của nó là hành động để đạt được phần thưởng cao nhất trong dài hạn. RL là một trong ba loại lớn của học máy, bao gồm: học không giám sát (unsupervised learning), học có giám sát (supervised learning) và học tăng cường. Học không giám sát được sử dụng để tìm các mẫu hoặc cấu trúc ẩn trong các tập dữ liệu chưa được phân loại hoặc dán nhãn. Bằng cách sử dụng học có giám sát, có thể huấn luyện máy tính áp dụng một nhãn cho một đầu vào nhất định. RL là một loại hình hoàn toàn khác biệt. Không giống như hai khung học tập kia vốn hoạt động bằng cách sử dụng tập dữ liệu tĩnh , RL làm việc với dữ liệu từ một môi trường động. Và mục tiêu không phải là phân cụm dữ liệu hay dán nhãn dữ liệu , mà là tìm ra chuỗi hành động tốt nhất sẽ tạo ra kết quả tối ưu. Cách RL giải quyết vấn đề, có thể được định nghĩa là một Quy trình Quyết định Markov (Markov Decision Process - MDP) , là bằng cách cho phép một tác nhân (agent) khám phá, tương tác và học hỏi từ môi trường. Bên trong tác nhân, có một hàm (chính sách - policy) nhận đầu vào là các quan sát trạng thái và ánh xạ chúng thành các hành động (outputs). Học tập là thuật ngữ dùng để chỉ quá trình điều chỉnh một cách có hệ thống các tham số trong hàm để hội tụ về chính sách tối ưu. RL sử dụng tín hiệu phần thưởng để đánh giá các cặp đầu vào-đầu ra và từ đó khám phá các đầu ra tối ưu cho mỗi đầu vào. Nói chung, năm lĩnh vực khác nhau cần được giải quyết với học tăng cường:

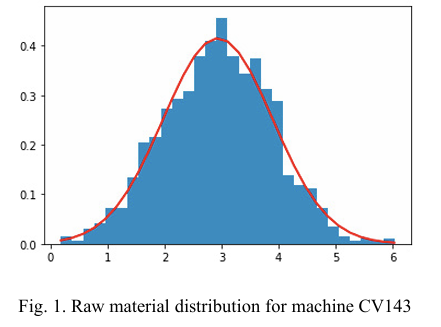
* Thiết lập môi trường
* Xác định phần thưởng dưới dạng một hàm của trạng thái và hành động
* Chọn cách để biểu diễn chính sách
* Chọn một thuật toán để huấn luyện tác nhân nhằm tìm ra các tham số chính sách tối ưu
* Khai thác chính sách bằng cách triển khai nó trong thực tế và xác minh kết quả.

Để tìm chính sách tối ưu, tối đa hóa phần thưởng kỳ vọng trong dài hạn, có nhiều phương pháp. Các thuật toán RL phổ biến nhất là phương pháp Monte Carlo, **Q-learning** và Deep Q-learning. Cụ thể, **Q-Learning** là một trong những thuật toán off-policy phổ biến với phương pháp khác biệt thời gian (temporal-difference approach) trong RL , và bao gồm một thuật toán độc lập mô hình (model-free algorithm) (tức là một phương pháp không sử dụng phân phối xác suất chuyển đổi và hàm phần thưởng liên quan đến MDP) , nhằm tìm hiểu chất lượng của hành vi để cho tác nhân biết nên thực hiện hành động nào trong điều kiện nào. Nó không cần một mô hình của môi trường , và có thể xử lý các chuyển đổi và phần thưởng ngẫu nhiên mà không cần điều chỉnh. Khung RL đơn giản và linh hoạt, do đó có thể áp dụng nó cho nhiều vấn đề khác nhau theo nhiều cách khác nhau.

# 3. Materials and Methods

## **3.1 Problem definition**

Trọng tâm của nghiên cứu này là quản lý tồn kho của **một sản phẩm đơn lẻ** tại một giai đoạn đơn lẻ của chuỗi cung ứng (tức là giai đoạn sản xuất). Cụ thể, vấn đề này bắt nguồn từ việc quan sát một **doanh nghiệp công nghiệp thực tế**. Chúng tôi đã xem xét bộ phận máy móc **CNC** (Computer Numerical Control), nơi các máy cần nguyên liệu thô để hoàn thành quá trình gia công. Để đáp ứng yêu cầu nguyên vật liệu thô từ các máy, đơn đặt hàng nguyên vật liệu thô được đặt tại kho nguyên vật liệu thô, được giả định có **công suất vô hạn**. Các công nhân vận hành máy kiểm tra mức tồn kho nguyên vật liệu thô mỗi **chu kỳ t** (mỗi ngày) và nhận nguyên vật liệu thô được yêu cầu sau một chu kỳ (tức là thời gian giao hàng - *procurement lead time*), được giả định là **hằng số và bằng một ngày** (nguyên vật liệu thô luôn được giao vào ngày làm việc tiếp theo). Phân tích này được thực hiện cho **máy móc quan trọng** (*critical machine*) và lượng tiêu thụ nguyên vật liệu thô là độc lập và được phân phối theo **phân phối chuẩn** (normal distribution) với **giá trị trung bình μ** và **độ lệch chuẩn σ**. Hình 1 cho thấy phân phối chuẩn của nguyên vật liệu thô cho máy móc quan trọng CV143.



Trình tự các sự kiện xảy ra như sau:

* Máy nhận lệnh sản xuất và dữ liệu về tồn kho trên máy được cập nhật.
* Nhu cầu nguyên vật liệu thô được thỏa mãn hoàn toàn thông qua tồn kho.
* Máy đặt hàng nguyên vật liệu thô theo chính sách bổ sung tồn kho của nó.

Trong nghiên cứu này, chính sách "đặt hàng lên đến" cổ điển, tức là **chính sách (r,q)**, được sử dụng. Chính sách này chủ yếu bao gồm việc rà soát liên tục mức tồn kho hiện có trên máy. Khi mức này nằm dưới một giá trị cố định r, một đơn hàng với số lượng q sẽ được đặt. Mục tiêu của nghiên cứu này là đánh giá **chính sách bổ sung tồn kho tối ưu** bằng cách sử dụng thuật toán **Q-learning**, dựa trên mô hình học tăng cường. Hơn nữa, **tổng chi phí**—bao gồm chi phí lưu trữ tồn kho (*holding cost*), chi phí đơn hàng tồn đọng (*backorder cost*) và chi phí phát hành đơn hàng (*ordering cost*)—được sử dụng làm thước đo hiệu suất.

Các chi phí được sử dụng trong mô hình được báo cáo trong Bảng 1 (vì lý do bảo mật của công ty, các con số được báo cáo không phải là số thực, nhưng chúng mang tính đại diện cho trường hợp). Cụ thể, chi phí của một đơn hàng đơn lẻ, **O**, bằng **50 €/đơn hàng**; chi phí lưu trữ là **10 €/chiếc-năm**; trong khi chi phí đơn hàng tồn đọng là **20 €/chiếc tồn đọng**. Thời gian cần thiết cho các hoạt động vận chuyển từ kho nguyên vật liệu thô đến máy móc quan trọng là **một ngày**.

Bảng 1. Các tham số chi phí được xem xét.

| Hạng mục Chi phí | Giá trị | Tham số |
| --- | --- | --- |
| Ordering cost (Chi phí đặt hàng) | 50 €/đơn hàng | O |
| Holding cost (Chi phí lưu trữ) | 10 €/chiếc-năm | h |
| Backorder cost (Chi phí tồn đọng) | 20 €/chiếc tồn đọng | b |

Số ngày làm việc mỗi năm là **365**. Hơn nữa, để triển khai các mô hình có xét đến **hàng tồn kho an toàn** (*safety stocks*), một **mức dịch vụ (SL)** bằng **95%** đã được xem xét, phù hợp với nhu cầu sản xuất của bộ phận máy **CNC** (Computer Numerical Control).

## **3.2 Reinforcement learning model**

Để áp dụng cơ chế **Học máy (Machine Learning)** vào vấn đề quản lý tồn kho và đặt hàng, cần phải xây dựng vấn đề dưới dạng một **Mô hình Học tăng cường (Reinforcement Learning - RL)**. Trong mục này, các yếu tố chính của mô hình RL được trình bày: tức là biến trạng thái (*state variable*), biến hành động (*action variable*) và hàm phần thưởng (*reward function*) cùng với thuật toán **Q-Learning** cho một chính sách quản lý tồn kho tối ưu nhằm bổ sung hàng cho các máy sản xuất.

### **3.2.1 State Variable**

quyết định cuối cùng trong các mô hình **Học tăng cường (RL)** được đưa ra dựa trên trạng thái của hệ thống, điều cần thiết là trạng thái của hệ thống phải cung cấp thông tin đầy đủ cho quá trình ra quyết định của các tác nhân học tập. Biến trạng thái được định nghĩa là cung cấp thông tin cần thiết cho việc ra quyết định của các tác nhân học tập. Có một trạng thái liên quan đến nghiên cứu này, được xác định bằng **S**. Véc-tơ S này được mô tả bởi tham số:

|  | (1) |
| --- | --- |

trong đó xác định **vị thế tồn kho** trên máy trong giai .

### **3.2.2 Action Variable**

**Biến hành động** thể hiện quyết định của tác nhân. Trong trường hợp cụ thể này, có hai hành động có thể:

* Quyết định **bổ sung** nguyên vật liệu thô được yêu cầu;
* Quyết định **KHÔNG bổ sung** nguyên vật liệu thô được yêu cầu.

Bằng cách xem xét các hành động trước đó, **vị thế tồn kho** (*inventory position*) thay đổi khi:

* Nhu cầu ​ đến, làm **giảm** vị thế tồn kho;
* Lô ​ (số lượng đặt hàng) đến, làm **tăng** vị thế tồn kho.

Cụ thể, bằng cách xem xét rằng **thời gian giao hàng (LT)** bằng một chu kỳ t, chúng ta có thể định nghĩa rằng trong trường hợp quyết định bổ sung:

|  | (2) |
| --- | --- |

trong khi, trong trường hợp quyết định là **KHÔNG bổ sung**:

|  | (3) |
| --- | --- |

Quyết định được chọn về cơ bản nằm giữa việc **đặt hàng hay không đặt hàng**.

### **3.2.3 Reward function**

Chính sách kiểm soát tồn kho được đề xuất nhằm mục đích **giảm thiểu tổng chi phí** của hệ thống (tức là tổng của chi phí lưu trữ, chi phí đơn hàng tồn đọng và chi phí đặt hàng). Do đó, hàm phần thưởng được xác định như sau:

|  | (4) |
| --- | --- |

rong đó rt​ chỉ ra **tổng chi phí** của tồn kho trong bước thời gian t, trong khi h, b và O lần lượt đại diện cho **chi phí lưu trữ đơn vị**, **chi phí đơn hàng tồn đọng đơn vị** và **chi phí phát hành một đơn hàng đơn lẻ**. Cụ thể, **chi phí lưu trữ** được tính khi vị thế tồn kho **dương** , trong khi **chi phí đơn hàng tồn đọng** phát sinh khi vị thế tồn kho **âm** .

### **3.2.4 Q-learning algorithm**

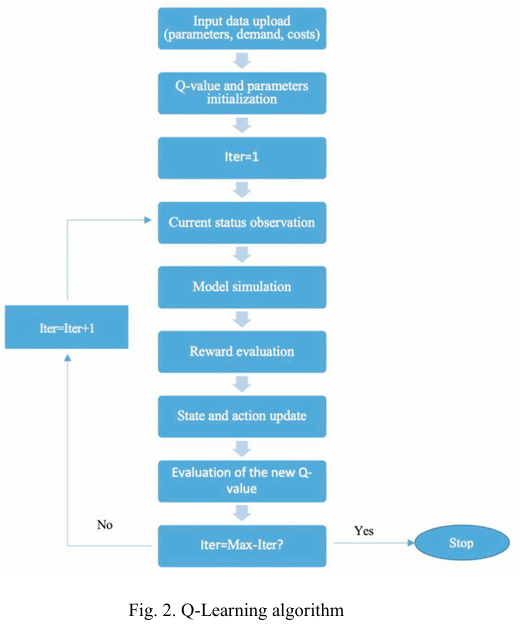
Thuật toán được thể hiện trong **Hình 2** dựa trên cơ chế **Q-learning**, là một phương pháp khác biệt thời gian (*time difference method*) được sử dụng để giải các mô hình Học tăng cường (RL). Thuật toán Q-learning được đề xuất để giải quyết vấn đề quản lý tồn kho. Học tăng cường (RL) tìm cách **tối đa hóa phần thưởng** bằng cách xác định số lượng đặt hàng thích hợp dựa trên kiến thức về trạng thái của tác nhân. Tuy nhiên, trong trường hợp cụ thể này, mục tiêu của kiểm soát tồn kho là **giảm thiểu chi phí** quản lý tồn kho. Do đó, hàm phần thưởng, đã được định nghĩa trước đó, nên được **giảm thiểu**.

Trong thuật toán, giá trị của các **hàm Q** đã được học trong quy trình lặp, ở cuối các thủ tục học tập (chúng tôi định nghĩa chu kỳ học tập bằng **1000 ngày**). Hành động tốt nhất trong mỗi trạng thái được chọn làm chính sách phân loại tối ưu để triển khai cho các trạng thái trong tương lai. Các giá trị hành động mới sau đó được cung cấp cho quá trình mô phỏng để **xác định lại các phần thưởng**, và những phần thưởng này được đưa ngược trở lại Học tăng cường (RL). Những chu kỳ này được lặp lại cho đến khi thuật toán học tăng cường **hội tụ** hoặc cho đến khi đạt đến **số lần lặp tối đa** (trong trường hợp này chúng tôi định nghĩa số lần lặp tối đa bằng **1000**).

Các giá trị của mỗi cặp **trạng thái-hành động** Q(s,a) chứa **phần thưởng tức thì** (*immediate reward*) mà tác nhân thu được khi thực hiện một hành động a trong trạng thái s. Những giá trị này được lưu trữ và cập nhật trong suốt quá trình học tập. Thật vậy, trong mỗi chu kỳ t, Q(​,​) được cập nhật theo phương trình sau:

| = | (5) |
| --- | --- |

Trong đó chỉ ra **tốc độ học tập** (*learning rate*) và kiểm soát mức độ quan trọng cần được đặt vào phần thưởng vừa trải nghiệm. Lưu ý rằng một quá nhỏ có thể khiến thuật toán **không hội tụ**, và một quá lớn (gần 1) cũng có thể làm tăng cường ảnh hưởng của một mẫu dữ liệu thiên lệch (*biased sample*) và khiến thuật toán không hội tụ. Thay vào đó, giá trị r đại diện cho **phần thưởng**. Cuối cùng, giá trị chỉ ra **hệ số chiết khấu** (*discount factor*), với 0<<1. Trong trường hợp của chúng tôi, chúng tôi định nghĩa =0.5 và =0.5.



Trong bảng sau, các biến chính của thuật toán được sử dụng

| Biến | Giải thích |
| --- | --- |
|  | Vị trí tồn kho trên máy trong chu kì t |
|  | Nhu cầu nguyên liệu thô trong chu kỳ t |
|  | Số lượng đặt hàng nguyên vật liệu thô đến vào chu kỳ t |
|  | Phần thưởng tại chu kỳ t |
|  | Phần thưởng được học tại chu kỳ t cho mỗi cặp (trạng thái, hành động) |

Ngôn ngữ lập trình **VBA** (*Visual Basic for Applications*) đã được sử dụng để triển khai **học tăng cường** (*reinforcement learning*).

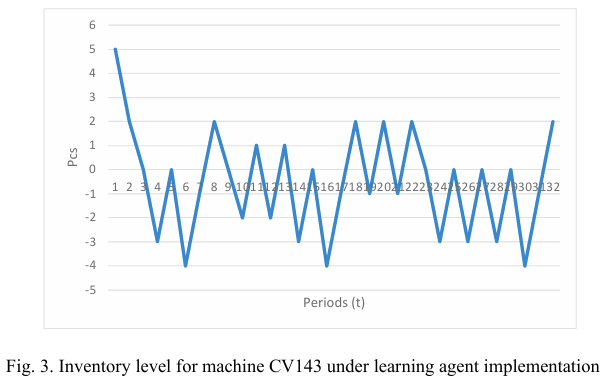
# 4. Results and discussion

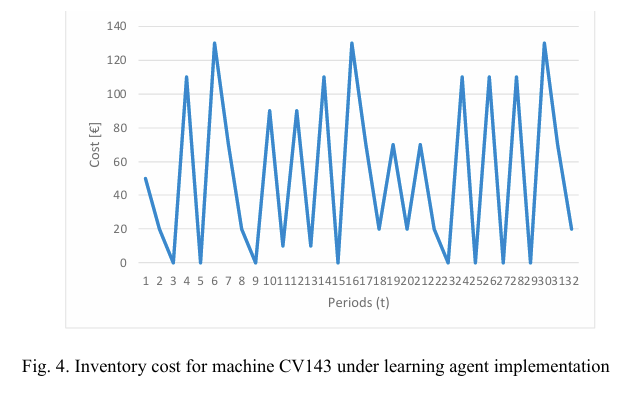
Các **kết quả mô phỏng** từ việc triển khai thuật toán **Q-learning** được thảo luận và so sánh với các kết quả thu được bằng cách triển khai các chính sách quản lý tồn kho **hiện tại** và **cổ điển**, với nhu cầu ngẫu nhiên (*stochastic demand*), cụ thể là:

* **Chính sách (r,q)**, bao gồm việc đặt lại một số lượng bằng q mỗi khi tồn kho vật lý giảm xuống dưới **điểm đặt hàng lại r**, kèm theo **hàng tồn kho an toàn** (*safety stocks*) để đối phó với sự không chắc chắn của nhu cầu trong suốt thời gian giao hàng (*lead time*).

Mô phỏng đã được thực hiện cho **máy móc quan trọng CV143**. Giá trị này được sử dụng để **huấn luyện mô hình Q-learning**. Việc triển khai thuật toán Q-learning cho quản lý tồn kho cho phép phân tích một năm làm việc với **nhu cầu biến đổi mỗi ngày**, nhờ đó tiệm cận gần hơn với tình hình thực tế. Cụ thể, nhu cầu nguyên vật liệu thô được coi là tuân theo **phân phối chuẩn** (*normal distributed*) với **giá trị trung bình =3 chiếc** và **độ lệch chuẩn =1 chiếc**. Kích thước lô hàng (*lot size*) được tính bằng cách sử dụng công thức **EOQ** (*Economic Order Quantity*) và giá trị bằng **6 chiếc**, trong khi r (điểm đặt hàng lại) bằng **3 chiếc**.

**Hình 3** cho thấy cả **xu hướng nhu cầu** (*demand trend*) và **mức tồn kho** (*level of stock*) được quản lý bởi tác nhân học tập cho máy gia công CV 143. **Hình 4** thể hiện **chi phí mỗi ngày** liên quan đến mức tồn kho, được xác định là tổng của chi phí lưu trữ (*holding*), chi phí đơn hàng tồn đọng (*backorder*) và chi phí đặt hàng (*ordering costs*).





Các **chi phí** thu được cho máy **CV143** đã được so sánh với các chi phí dưới **chính sách tồn kho (r,q)** đã giới thiệu trước đó. Như có thể thấy từ **Bảng 3**, thuật toán **Q-Learning** dẫn đến chi phí thấp nhất, ngay cả khi mô phỏng này dựa trên các giá trị biến đổi cho nhu cầu. So với chính sách (r,q) truyền thống, như được thể hiện trong Bảng 3, phương pháp **Q-learning cho phép giảm tổng chi phí** (*reduce the overall costs*).

Bảng 3. So sánh tổng chi phí

| Nhu cầu Trung bình [pcs] | Độ Lệch Chuẩn Nhu cầu [pcs] | Hệ số biến thiên | Tổng chi phí (chính sách r,Q) | Tổng chi phí (Q-Learning) |  |
| --- | --- | --- | --- | --- | --- |
| 2 | 1 | 0,5 | 16.391,47€ | 14.348,15€ | -12,47% |
| 3 | 1 | 0,333 | 20.025,05€ | 19.012,85€ | -5,05% |
| 5 | 1 | 0,2 | 25.902,31€ | 23.980,50€ | -7,42% |

Cụ thể, có thể nhận thấy rằng khi **tăng sự biến thiên** (*variability*), phương pháp **Q-learning** thể hiện **hiệu suất vượt trội** hơn so với phương pháp truyền thống. Tiếp theo, chúng tôi báo cáo sự so sánh giữa **Tổng chi phí** của phương pháp Q-learning và phương pháp truyền thống bằng cách **tăng thành phần chi phí đơn hàng tồn đọng** (*backorder cost*) cho kịch bản với nhu cầu bằng **3 chiếc**. **Bảng 4** cho thấy phương pháp Q-learning **vượt trội hơn** phương pháp truyền thống khi chi phí đơn hàng tồn đọng tăng lên. Điều này là do khả năng của phương pháp Q-learning trong việc **điều chỉnh quyết định mua** ở các mức vị thế tồn kho khác nhau, trong khi phương pháp truyền thống chỉ mua khi vị thế tồn kho nằm dưới **điểm đặt hàng lại** (*reorder point*).

Bảng 4. So sánh tổng chi phí với sự biến đổi của thành phần chi phí đơn hàng tồn đọng

| Nhu cầu Trung bình [pcs] | Độ Lệch Chuẩn Nhu cầu [pcs] | Hệ số biến thiên | Chi phí đơn hàng tồn đọng [€/chiếc tồn đọng] | Chi phí lưu trữ [€/chiếc-năm] | Tổng chi phí (chính sách r,Q) | Tổng chi phí (Q-Learning) | D% |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 1 | 0,333 | 20 | 10 | 20.025,05 € | 19.012,85 € | -5,05% |
| 3 | 1 | 0,333 | 60 | 10 | 25.805,50 € | 20.813,52 € | -19,34% |
| 3 | 1 | 0,333 | 90 | 10 | 29.505,38 € | 23.821,12 € | -19,27% |
| 3 | 1 | 0,333 | 120 | 10 | 33.959,60 € | 26.412,62 € | -22,22% |

Ngoài các lợi ích về kinh tế, học tăng cường còn thể hiện một số **lợi thế** khi so sánh với các chính sách quản lý tồn kho khác.

Ưu điểm đầu tiên là các mô hình **RL** có thể được huấn luyện để xem xét **khối lượng lớn dữ liệu**, cho phép chúng khám phá các xu hướng và mẫu hình cụ thể mà đôi khi không rõ ràng; ví dụ, nó có khả năng xác định một mối quan hệ ngẫu nhiên giữa hai sự kiện. Hơn nữa, nó **không nhất thiết cần sự can thiệp của con người**, vì khả năng học hỏi được chuyển giao cho máy móc, cho phép chúng đưa ra dự đoán và cải tiến các thuật toán. Một ưu điểm mạnh mẽ khác của RL là sự **tiến bộ liên tục**, tức là khi thuật toán tích lũy kinh nghiệm, nó tối ưu hóa độ chính xác và hiệu quả, nhờ đó đưa ra các quyết định ngày càng tốt hơn.

Từ nghiên cứu điển hình này, Học tăng cường được triển khai trong các chính sách cung ứng nguyên vật liệu thô sẽ cho phép **quản lý dữ liệu tối ưu** và cải thiện rõ rệt việc xử lý vật liệu. Hơn nữa, nó sẽ không yêu cầu nỗ lực của con người, mà mọi thứ sẽ được quản lý bởi **trí tuệ nhân tạo**, do đó giảm đáng kể thời gian dành cho việc kiểm tra và quản lý vật liệu trên máy. Trên thực tế, hiện nay công ty vẫn đang dựa kỹ năng quản lý của mình vào các kỹ năng và kiến thức được phát triển qua nhiều năm của nhân viên.

Tuy nhiên, công nghệ này đòi hỏi một **tập hợp dữ liệu khổng lồ**—đôi khi khó có được—để huấn luyện mô hình, và những dữ liệu này phải có **chất lượng tuyệt vời**. Hơn nữa, một khi dữ liệu đã được thu thập, Học tăng cường cần **đủ thời gian** để thuật toán Q-learning học và huấn luyện dữ liệu, nhằm đạt được mục đích với độ chính xác đáng kể.

Ngoài ra, học máy là **tự chủ** (*autonomous*) nhưng lại **rất dễ bị lỗi** nếu thuật toán được huấn luyện với lượng dữ liệu không đủ. Điều này có thể dẫn đến các dự đoán không chính xác từ một tập huấn luyện **thiên lệch** (*biased training set*).

# 5. Conclusions and future research

Trong nghiên cứu này, chúng tôi giới thiệu một ứng dụng của **Học máy (Machine Learning)**, dựa trên **Học tăng cường (Reinforcement Learning)**, có khả năng tối ưu hóa các tham số của chính sách tồn kho. Cụ thể, phần thưởng của mô hình được định nghĩa bằng cách xem xét **chi phí lưu trữ** (*holding costs*), **chi phí đơn hàng tồn đọng** (*backorder cost*) và **chi phí đặt hàng** (*ordering cost*).

Mô hình được áp dụng cho một **trường hợp công nghiệp thực tế**, và các kết quả của mô hình **Q-learning** cho thấy **sự giảm thiểu đáng kể chi phí tồn kho** so với chính sách tồn kho truyền thống (r,q).

Để phát triển nghiên cứu này hơn nữa, có thể thực hiện các mô phỏng bằng cách giới thiệu **thời gian giao hàng ngẫu nhiên** (*stochastic lead time*) và/hoặc tối ưu hóa **kích thước lô hàng Q** thông qua Q-learning.