**UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS**

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMATICA 

**CURSO:**

[AUTOMATIZACIÓN Y CONTROL DE SOFTWARE](https://classroom.google.com/u/1/c/NzA3Njc4MTg0Mzk2)

**TEMA:**

Implementación del algoritmo Q-learning para el caso del pendulo invertido

**PROFESOR:**

Yessica Rosas Cueva

**ESTUDIANTES:**

* Chuquispuma Merino, Fabricio Vidal: [0009-0008-6474-2691](https://orcid.org/0009-0008-6474-2691)
* Felix Huayhua, Axel Patrick: [0009-0009-6357-9308](https://orcid.org/0009-0009-6357-9308)
* Fernandez Camacho, Geomar Willy: [0009-0005-0838-6504](https://orcid.org/0009-0005-0838-6504)
* Saavedra Monterrey Max Bruno: [0009-0003-7066-2499](https://orcid.org/0009-0003-7066-2499)
* Sanchez Saldaña, Sergio Antonio: [0009-0000-8297-0761](https://orcid.org/0009-0000-8297-0761)

**Lima - Perú**

**2024**

**Resumen**

La implementación de Q-Learning para el control del péndulo invertido representa un caso de estudio significativo en el ámbito del aprendizaje por refuerzo. Este artículo presenta la descripción del problema y su implementación en Python con el fin de desarrollar un agente capaz de mantener el equilibrio de un péndulo invertido sobre un carro móvil utilizando algoritmos de Q-Learning.

Se describe la configuración del entorno de simulación, que se realiza mediante la biblioteca pygame, permitiendo la visualización y la interacción en tiempo real. El agente se entrena a través de una política epsilon-greedy, donde la exploración y explotación de acciones se balancea mediante un parámetro epsilon (ϵ), que disminuye progresivamente a medida que avanza el entrenamiento. La tabla Q, que almacena las estimaciones de valor de acción-estado, se actualiza utilizando la ecuación de Bellman, integrando recompensas inmediatas y el valor esperado de acciones futuras, regulado por la tasa de aprendizaje y la tasa de descuento.

Los resultados obtenidos muestran las distintas pruebas que se realizaron modificando los parametros, tras un número significativo de episodios de entrenamiento, el agente logra equilibrar con éxito el péndulo invertido, evidenciado por las recompensas acumuladas y la estabilidad del sistema en condiciones de prueba. Este estudio no solo valida la efectividad del Q-Learning en problemas de control dinámico, sino que también ofrece una base para futuras investigaciones en el desarrollo de algoritmos de aprendizaje por refuerzo aplicados a sistemas más complejos.

Palabras clave: Q-learning, péndulo invertido, aprendizaje reforzado, implementación, estados, acciones

**Contenido**

[**Resumen** 2](#_Toc184733703)

[Capítulo 1: Introducción 4](#_Toc184733704)

[Capítulo 2: Marco teórico 6](#_Toc184733705)

[Capítulo 3: Estado del arte 8](#_Toc184733706)

[Artículo 1 8](#_Toc184733707)

[Artículo 2 9](#_Toc184733708)

[Artículo 3 11](#_Toc184733709)

[Artículo 4 12](#_Toc184733710)

[Artículo 5 12](#_Toc184733711)

[Capítulo 3: Metodología 15](#_Toc184733712)

[Metodología del desarrollo 15](#_Toc184733713)

[Formulación del Problema 15](#_Toc184733714)

[Implementación del Entorno 16](#_Toc184733715)

[Capítulo 4: Desarrollo 19](#_Toc184733716)

[Modelado del Sistema 19](#_Toc184733717)

[Sistema de Recompensas 19](#_Toc184733718)

[Algoritmo Q-Learning 20](#_Toc184733719)

[Estados y Discretización 20](#_Toc184733720)

[Acciones 21](#_Toc184733721)

[Políticas 21](#_Toc184733722)

[Selección de Hiperparámetros 21](#_Toc184733723)

[Parámetros del Entorno y Fuerza 22](#_Toc184733724)

[Exploración y Explotación 22](#_Toc184733725)

[Capítulo 5: Validación y resultados 23](#_Toc184733726)

[Capítulo 6: Conclusiones y Discusiones 35](#_Toc184733727)

[Referencias 38](#_Toc184733728)

Capítulo 1: Introducción

**Contexto del problema**

El sistema de péndulo invertido es un clásico en el estudio de control de sistemas no lineales, caracterizado por su inestabilidad intrínseca. Este desafío implica estabilizar un péndulo en posición vertical mediante la aplicación de fuerzas horizontales en un carro móvil. La interacción entre la dinámica del carro, las fuerzas gravitatorias y el movimiento angular del péndulo genera un sistema altamente no lineal.

La modelización del péndulo invertido tradicionalmente se realiza a través de la segunda ley de Newton, lo cual implica resolver ecuaciones diferenciales no lineales que describen el movimiento del carro, la posición angular del péndulo y las fuerzas aplicadas. Sin embargo, este enfoque puede resultar complejo debido a la inestabilidad inherente del sistema y la necesidad de técnicas avanzadas como la linealización o el uso de controladores como PD, compensadores de estado o estrategias basadas en aprendizaje por refuerzo (Ahmad Ataka et al., 2023).

En este caso, se opta por implementar Q-Learning, un método de aprendizaje por refuerzo que permite al sistema aprender estrategias de control a través de iteraciones con su entorno, sin depender de un modelo matemático explícito. Este enfoque conceptualiza el problema como un Proceso de Decisión de Markov (MDP), donde un agente aprende a maximizar recompensas acumulativas al estabilizar el péndulo. A través de una tabla Q, que se actualiza en función de las recompensas obtenidas para cada estado y acción, el agente desarrolla una política óptima tras un entrenamiento adecuado. Este enfoque ha demostrado su efectividad incluso frente a las dificultades de la discretización del espacio de estados y las limitaciones computacionales (Sardor Israilov et al., 2023).

El péndulo invertido no solo representa un problema académico y práctico en el campo del control y la inteligencia artificial, sino que también sirve como una valiosa herramienta de aprendizaje para explorar dinámicas no lineales, algoritmos de aprendizaje reforzado y estrategias de estabilización en sistemas inestables.

**Formulación del problema**

La estabilización del péndulo invertido, un sistema dinámico no lineal e inestable, exige estrategias avanzadas para mantener su posición vertical mientras se minimiza el desplazamiento del vehículo. Esta investigación propone implementar el algoritmo de aprendizaje por refuerzo Q-learning, evaluando parámetros como la tasa de aprendizaje y el factor de descuento para optimizar la estabilidad del sistema. Mediante simulaciones y experimentos, se busca validar la eficacia de esta técnica, destacando su capacidad para abordar sistemas complejos sin requerir un modelo matemático explícito (Dev et al., 2024; Sardor Israilov et al., 2023).

**Objetivo general**

Implementar un sistema de control basado en aprendizaje por refuerzo mediante el algoritmo Q-learning para estabilizar un péndulo invertido modificado, considerando las limitaciones físicas y la complejidad del sistema.

**Objetivos específicos**

* Investigar los fundamentos teóricos del aprendizaje por refuerzo y su aplicación en sistemas dinámicos no lineales.
* Diseñar e implementar el algoritmo Q-learning para abordar el control del péndulo invertido.
* Optimizar los parámetros del algoritmo Q-learning, como la tasa de aprendizaje y el factor de descuento, para mejorar el desempeño del sistema.
* Validar los resultados del aprendizaje mediante simulaciones y análisis cuantitativo de métricas clave como estabilidad y desplazamiento del vehículo.

Capítulo 2: Marco teórico

**Péndulo Invertido:**

El péndulo invertido, un sistema dinámico de naturaleza no lineal y de equilibrio intrínsecamente inestable, ha sido empleado como caso de estudio para el análisis y desarrollo de diversos métodos controladores. Este sistema implica la disposición de un péndulo sobre un carrito móvil, cuyo propósito es mantener la barra en una posición vertical mientras se estabiliza el movimiento del carrito. Para Ahmad et al., (2023) este objetivo se alcanza mediante la implementación de una fuerza regulada sobre el carrito, que compensa las desviaciones angulares del péndulo.

**Aprendizaje por Refuerzo:**

Dentro del campo del aprendizaje automático, el aprendizaje por refuerzo (RL) se destaca como un enfoque que pone énfasis en la comunicación entre un agente y su entorno, con la meta de aumentar las recompensas obtenidas a lo largo del tiempo. Contrario a otras técnicas, RL no necesita modelos detallados del sistema, lo que lo convierte en una herramienta especialmente beneficiosa en escenarios donde los parámetros del sistema son desconocidos o complicados de representar (Israilov et al., 2023).

La base conceptual del aprendizaje por refuerzo se sustenta en el Proceso de Decisión de Markov (MDP), el cual engloba estados, acciones, recompensas y políticas de decisión. Mediante este esquema, el agente puede adquirir una estrategia ideal mediante un proceso de experimentación y corrección. Para Dev, Chowdhury y Schoen (2024), uno de los métodos más populares en RL es el algoritmo Q-learning, que de forma iterativa ajusta una tabla Q para estimar los valores esperados de recompensa de las acciones en distintos estados.

Asimismo, gracias a técnicas innovadoras como las Redes Neuronales Profundas de Q-Learning (DQN), se han logrado sobrepasar las restricciones que implicaba la discretización de los estados y acciones en el algoritmo convencional. Israilov et al. (2023) señala que las redes neuronales funcionan como estimadores de funciones, brindando una mayor versatilidad y eficacia en la resolución de desafíos de control complicados, como el péndulo invertido.

Además, RL ha mostrado su fortaleza al manejar situaciones complejas y sistemas inestables con éxito. Esto abarca el uso en situaciones reales donde las circunstancias son naturalmente impredecibles, como la interferencia en las mediciones y la resistencia desigual. Investigaciones recientes han demostrado que incluso modificaciones sencillas en el algoritmo de Q-learning, como la incorporación de funciones de recompensa más elaboradas, tienen la capacidad de aumentar de manera considerable la estabilidad y eficacia del control en dispositivos como el péndulo invertido (Ahmad et al., 2023).

**Q-Learning:**

El algoritmo Q-learning es una estrategia de aprendizaje por refuerzo sin modelo que ha probado ser eficaz en actividades de control complicadas, como la tarea de mantener estable un péndulo invertido. Este enfoque se basa en adquirir una estrategia óptima a través de la interacción con el entorno, analizando las recompensas vinculadas a diversas acciones y situaciones. La función de valor Q, que indica la utilidad esperada de llevar a cabo una acción particular en una situación dada, se modifica de forma iterativa mediante la aplicación de la ecuación de Bellman. Esto posibilita que el agente mejore sus estrategias con el transcurso del tiempo (Dev, Chowdhury y Schoen, 2024).

Dentro del ámbito del péndulo invertido sobre un carro móvil, el algoritmo de aprendizaje por refuerzo Q-learning ha sido utilizado de manera exitosa para abordar la cuestión del equilibrio y la oscilación hacia arriba.

Capítulo 3: Estado del arte

## Artículo 1

**Título:** Aprendizaje por refuerzo profundo difuso guiado por conocimiento

**Aporte:** El artículo propone un nuevo método de aprendizaje por refuerzo profundo, llamado KFDQN, que utiliza sistemas difusos para integrar conocimiento previo en el proceso de aprendizaje. Esto mejora la eficiencia del aprendizaje, la estabilidad del entrenamiento y el rendimiento general del modelo. KFDQN integra el conocimiento en la selección de acciones y en la definición de objetivos de aprendizaje. Se valida el algoritmo mediante simulaciones y experimentos con robots móviles, mostrando una mejora significativa en eficiencia y rendimiento. El estudio destaca la importancia de combinar el aprendizaje automático con el conocimiento humano para lograr mejores resultados.

**Proceso:** El estudio se basa en la combinación de aprendizaje por refuerzo profundo (DQN) con un sistema difuso para representar el conocimiento. El método HYAS (Hybrid Action Selection Strategy) se usa para integrar el conocimiento en la selección de acciones, reduciendo la aleatoriedad. HYL (Hybrid Learning Method) combina conocimiento y aprendizaje automático en el objetivo de aprendizaje. Se utilizan tareas en OpenAI Gym (CartPole, MountainCar) y en robots móviles (llegar a un objetivo y evitar obstáculos) para validar el método. Se evalúa el rendimiento en términos de eficiencia de aprendizaje, estabilidad y rendimiento del modelo.

**Resultados:** Los resultados experimentales demuestran una mejora significativa en la eficiencia del aprendizaje (28.6%), el rendimiento (19.56%) y la estabilidad del entrenamiento de KFDQN en comparación con los algoritmos base. KFDQN mostró una mejor eficiencia de aprendizaje en tareas de planificación a corto y largo plazo. En robots móviles, KFDQN logró mejor rendimiento en tareas de alcanzar un objetivo y esquivar obstáculos. KFDQN superó a algoritmos base en estabilidad y tiempos de ejecución en ambas tareas. El método propuesto destaca la eficacia de la integración del conocimiento previo en el aprendizaje por refuerzo.

## Artículo 2

**Título:** Control de péndulo invertido: un estudio comparativo entre el control convencional y el aprendizaje por refuerzo

**Aporte:** Este artículo se aborda la comparación entre métodos de control convencionales (compensador basado en locus de raíces, compensador de estado y control proporcional-derivativo, PD) y un enfoque basado en aprendizaje por refuerzo (Proximal Policy Optimization, PPO) para controlar un sistema de péndulo invertido. Su contribución principal radica en evidenciar que, bajo un entrenamiento adecuado, el aprendizaje por refuerzo puede lograr un rendimiento comparable a los métodos tradicionales sin requerir conocimiento previo del modelo del sistema. Esto resulta valioso en sistemas donde el modelado es complejo o inaccesible.

**Proceso:** El procedimiento descrito en el artículo se dividió en tres fases principales. En el primer apartado, se trató el proceso de modelado y formulación del sistema de péndulo invertido. Mediante los principios de las leyes de Newton, se ha desarrollado un modelo dinámico que explica las relaciones entre el vehículo y el péndulo. Posteriormente, se procedió a linealizar este modelo no lineal con el fin de simplificar la implementación de técnicas de control convencionales. Para la metodología basada en el aprendizaje por refuerzo, se creó una función de recompensa que tuvo en cuenta la ubicación y la inclinación del péndulo, con el objetivo de mantenerlo en una posición estable.  
En la fase dos, se llevaron a cabo la implementación de los algoritmos de control. Se han elegido tres métodos de control convencionales: el compensador basado en el lugar de las raíces, el compensador de estado y el control proporcional-derivativo (PD). Además, se empleó el algoritmo de optimización de políticas proximales (PPO) como un ejemplo de los enfoques de aprendizaje por refuerzo. El último modelo fue entrenado con 250,000 episodios en un entorno de simulación modificado de OpenAI Gym. Este entorno es compatible con acciones continuas y proporciona recompensas basadas en el comportamiento del sistema.  
Por último, en la tercera fase, se procedió a evaluar los algoritmos a través de pruebas realizadas en un modelo de simulación del péndulo invertido. Se examinaron las respuestas temporales, abarcando el tiempo máximo, el tiempo de estabilización y el exceso de respuesta, además de las respuestas en equilibrio, teniendo en cuenta el error en equilibrio y la energía total consumida.

**Resultados:** El examen de los resultados indicó que el compensador que se basa en el locus de raíces tuvo el rendimiento más sobresaliente en general, presentando un tiempo de estabilización rápido y un error en estado estacionario mínimo. Este procedimiento consiguió estabilizar el péndulo de forma eficaz, destacándose sobre otros enfoques en diversos aspectos fundamentales. El algoritmo PPO ha mostrado ser competitivo, especialmente en cuanto a eficiencia energética, al requerir menos energía para estabilizar el sistema en comparación con las técnicas convencionales. No obstante, el rendimiento durante el periodo de estabilización y el momento de mayor actividad fue deficiente, esto se debió a las restricciones de la función de recompensa empleada, la cual no incluyó de manera explícita criterios de rendimiento temporal. Por otra parte, los métodos de compensación de estado y control proporcional-derivativo mostraron un rendimiento más restringido. Específicamente, el control proporcional-derivativo presentó oscilaciones continuas y no logró estabilizar adecuadamente el péndulo en el tiempo asignado para las pruebas.

## Artículo 3

**Titulo:** Modeling an Inverted Pendulum via Differential Equations and Reinforcement Learning Techniques

**Aporte:** Este trabajo demuestra la viabilidad de usar técnicas de aprendizaje por refuerzo (RL), específicamente Deep Q-Network (DQN), para controlar un péndulo invertido sin recurrir a los métodos de control clásicos como PID o LQR. El aporte radica en la aplicación exitosa de DQN a un problema de control dinámico no lineal, mostrando una alternativa robusta y adaptable a la complejidad del sistema. El documento también proporciona un desarrollo detallado de las ecuaciones diferenciales que modelan el sistema y una explicación clara de los fundamentos del aprendizaje por refuerzo.

**Proceso:** El proceso de investigación se puede dividir en estas etapas:  Desarrollo de las ecuaciones diferenciales que describen la dinámica del péndulo invertido, considerando tanto el modelo lineal como el no lineal.Aplicación de la técnica de Deep Q-Learning (DQN) para resolver el problema de control. Esto implica definir el espacio de estados, las acciones posibles, la función de recompensa y la arquitectura de la red neuronal para aproximar la función Q. Entrenamiento de la red neuronal DQN utilizando un entorno de simulación para el péndulo invertido. Se describe la fase de aprendizaje del sistema, incluyendo la optimización de la función de pérdida y la convergencia del algoritmo. Presente los resultados del entrenamiento en forma de gráficas que ilustran la convergencia del algoritmo y su rendimiento en la tarea de estabilizar el péndulo.

**Resultados:**  Los resultados muestran que el algoritmo DQN entrenado logra controlar con éxito el péndulo invertido, manteniéndolo en equilibrio durante un período prolongado. Se presenta una gráfica que muestra la reducción en el número de pasos necesarios para estabilizar el péndulo a medida que avanza el entrenamiento. Aunque no se presentan métricas cuantitativas precisas de rendimiento (tiempo de respuesta, sobreimpulso, etc.), la gráfica indica claramente la efectividad del método propuesto en comparación con los métodos de control tradicionales. El éxito del DQN en este problema no lineal y complejo valida su aplicabilidad como una alternativa a los métodos clásicos.

## Artículo 4

**Titulo:** Control basado en Q-learning para balanceo y equilibrio de péndulo invertido

**Aporte:** La investigación vincula la teoría de control clásica con técnicas

De aprendizaje por refuerzo, contribuyendo para demostrar versatilidad y potencial del aprendizaje automático en tareas de control. Se identificó que el factor de descuento óptimo influye significativamente en el resultado del aprendizaje y en la estabilidad del sistema.

**Proceso:** Para la simulación del modelo se empleó una representación precisa de la dinámica física la cual rigue al sistema, combinada con un algoritmo de aprendizaje por refuerzo para el control de movimientos del carro, las fuerzas actuantes son la fuerza aplicada la fuerza gravitacional del pendulo y la inercia del pendulo. Se establecio un limite en cada lado con una distancia igual desde la posición inicial del carro, lo que se busca es que el angulo en el que oscila el pendulo este dentro de los 20 grados. Con relación al estado, el estado siguiente responde al estado actual y a la acción realizada, los parámetros actualizados derivan de ecuaciones físicas simplificadas. Se hicieron usos de 3 tasas de aprendizajes (0.5,0.7 y 0.9) y 3 factores de descuentos (0.1,0.2 y 0.5)

**Resultados:** El factor de aprendizaje de 0,7 alcanzo la estabilidad por un periodo de (3,6 s), lo que implica que es el equilibrio sostenido mas largo entre los identificados. Con relación al factor de descuento el mas destacable es el 0,2 , es el que establece el equilibrio mas optimo

## Artículo 5

**Título:** Reinforcement learning approach to control an inverted pendulum: A general framework for educational purposes

**Aporte:** Este artículo presenta un marco general para el control de un péndulo invertido utilizando algoritmos de aprendizaje por refuerzo (Reinforcement Learning, RL), destacándose por su aplicabilidad en entornos educativos. Proporciona una guía detallada para implementar los métodos Q-Learning y Deep Q-Networks (DQN) tanto en simulaciones como en experimentos reales. Además, aborda desafíos clave, como el ruido y la fricción, a través de simulaciones que replican las condiciones del mundo real, y valida la robustez de estos métodos para condiciones iniciales variadas. Al incluir un repositorio de código abierto, el trabajo fomenta la experimentación y el aprendizaje práctico de tecnologías avanzadas de control en un contexto académico, abriendo oportunidades para explorar el aprendizaje automático en aplicaciones dinámicas complejas.

**Proceso:** El estudio combina simulaciones y experimentos físicos para implementar el control del péndulo invertido mediante RL. Primero, se desarrolló un modelo físico del sistema y se definió un entorno RL que utiliza estados, acciones y recompensas para guiar el aprendizaje del agente. Se implementaron algoritmos Q-Learning y DQN, comparándolos en términos de eficiencia y precisión. Las simulaciones incluyeron variaciones en parámetros físicos clave como el voltaje, la fricción y el ruido, para evaluar el impacto en el aprendizaje y la estabilidad del control. Finalmente, se realizaron experimentos reales utilizando un sistema equipado con sensores y un motor controlado por un microcontrolador, validando los resultados obtenidos en simulaciones y destacando las ventajas de DQN frente a Q-Learning.

**Resultados:** Los resultados muestran que el algoritmo DQN logró estabilizar el péndulo invertido de manera eficiente y robusta frente a diversas condiciones iniciales y parámetros físicos, mientras que Q-Learning presentó limitaciones significativas en cuanto a tiempo de aprendizaje y precisión. En los experimentos, se identificó que un voltaje mínimo de 5.9 V es esencial para alcanzar un control óptimo, y las simulaciones confirmaron la efectividad del enfoque para manejar ruido y fricción moderados. Este trabajo destaca la capacidad de DQN para generalizar y adaptarse a condiciones variables, ofreciendo un marco educativo sólido para explorar aplicaciones de aprendizaje por refuerzo en sistemas dinámicos.

Capítulo 3: Metodología

## Metodología del desarrollo

La metodología seguida para resolver el problema del péndulo invertido con un carro deslizante basado en Q-Learning se dividió en varias fases clave. Estas fases incluyen la formulación del problema, la implementación del entorno, el diseño del agente de aprendizaje, y la evaluación del desempeño del modelo. A continuación, se describe cada una de estas etapas en detalle.

### Formulación del Problema

El objetivo principal de este estudio fue desarrollar un sistema de control para un péndulo invertido sobre un carro deslizante, que mantuviera el péndulo en su posición vertical y el carro lo más cerca posible de su posición central. Se definieron los siguientes aspectos clave del problema:

* **Espacio de estados**: El estado del sistema fue representado mediante cinco variables continuas: el ángulo del péndulo (), la velocidad angular (), la posición del carro (), la velocidad del carro (), y la aceleración del carro (). Para facilitar el uso del algoritmo de Q-Learning, estas variables fueron discretizadas en un conjunto finito de valores.
* **Acciones posibles**: El agente podía ejecutar tres posibles acciones: mover el carro hacia la izquierda, mantener el carro quieto, o moverlo hacia la derecha. Cada acción tiene un impacto directo en el sistema y puede influir en la estabilidad del péndulo y la posición del carro.
* **Función de recompensa**: La función de recompensa fue diseñada para penalizar los estados donde el péndulo no se encuentra en posición vertical o cuando el carro se aleja demasiado del centro del riel. La recompensa se calcula en función de la distancia del ángulo al valor cero (posición vertical) y de la posición del carro respecto al centro del riel.

## Implementación del Entorno

El entorno fue modelado siguiendo las leyes físicas que gobiernan el movimiento del péndulo y el carro. Se utilizó un modelo simplificado que describía el comportamiento dinámico del sistema, donde las fuerzas que actuaban sobre el carro y el péndulo se reflejaban en las variables de estado mencionadas previamente.

* **Discretización del espacio de estados**: Debido a que el Q-Learning requiere un espacio de estados finito, las variables continuas del sistema (por ejemplo, , , etc.) fueron discretizadas. Se definieron rangos específicos para cada variable y se estableció una resolución de discretización adecuada para garantizar que el espacio de búsqueda fuera lo suficientemente denso para el aprendizaje, pero sin ser demasiado grande como para hacerlo computacionalmente ineficiente.
* **Dinámica del sistema**: La evolución del sistema tras la aplicación de una acción se modeló a través de ecuaciones diferenciales que describen el movimiento del péndulo y el carro. Estas ecuaciones fueron resueltas de forma iterativa en cada paso del agente.

#### Implementación del Agente de Q-Learning

El agente de aprendizaje fue diseñado utilizando el algoritmo Q-Learning, un enfoque de refuerzo basado en la actualización de una tabla de valores , que representa la calidad de tomar una acción en un estado .

* **Estrategia de exploración-explotación**: Se utilizó una política *epsilon-greedy* para gestionar la exploración y explotación. Al principio, el agente realizó más exploración, eligiendo acciones de manera aleatoria con una probabilidad de para descubrir el entorno. A medida que el aprendizaje avanzaba, el valor de disminuía gradualmente, favoreciendo la explotación de las acciones con los valores más altos.
* **Actualización de la tabla** : En cada iteración, después de realizar una acción y observar el siguiente estado y la recompensa obtenida, el valor de la tabla se actualizaba utilizando la ecuación de Bellman. Esta actualización permitió al agente aprender gradualmente cuál era la mejor acción para tomar en cada estado para maximizar la recompensa acumulada.

#### Entrenamiento y Evaluación

El agente fue entrenado mediante simulaciones repetidas del sistema. Durante cada episodio de entrenamiento, el agente interactuó con el entorno, seleccionó acciones según su política , y actualizó la tabla en función de las recompensas obtenidas. Se definieron los siguientes criterios para la evaluación del desempeño:

* **Número de episodios de entrenamiento**: El agente fue entrenado durante un número fijo de episodios para asegurarse de que el proceso de aprendizaje fuera lo suficientemente largo para que el agente pudiera converger hacia una política efectiva.
* **Desempeño del agente**: El desempeño del agente se evaluó observando la capacidad del sistema para mantener el péndulo en posición vertical y el carro centrado durante los episodios de prueba. La recompensa promedio obtenida por el agente a lo largo de los episodios de prueba se utilizó como una métrica para medir la calidad de la política aprendida.

#### Ajuste de Hiperparámetros

Durante el entrenamiento, se realizaron ajustes en varios hiperparámetros clave del algoritmo para optimizar el proceso de aprendizaje:

* **Tasa de aprendizaje** : Controla la velocidad con la que se ajustan los valores en la tabla en función de las recompensas obtenidas. Se utilizó un valor moderado para evitar cambios bruscos en la política durante el entrenamiento.
* **Factor de descuento** : Este parámetro determina la importancia de las recompensas futuras frente a las recompensas inmediatas. Se fijó en , lo que indica que las recompensas futuras son casi tan importantes como las inmediatas.
* **Tasa de exploración inicial** : Inicialmente, se configuró en , favoreciendo la exploración. Posteriormente, se redujo gradualmente hasta un valor mínimo de , favoreciendo la explotación a medida que el agente aprendía.

Capítulo 4: Desarrollo

## Modelado del Sistema

El sistema físico está compuesto por un carro de masa ​ que puede desplazarse sobre un riel horizontal y un péndulo de masa y longitud fijado en la parte superior del carro. La dinámica de este sistema se describe mediante ecuaciones diferenciales acopladas que gobiernan el movimiento del carro y la oscilación del péndulo (Sutton & Barto, 2018). Las ecuaciones del sistema son las siguientes:

1. **Ecuación del péndulo:**
2. **Ecuación del carro:**

**Donde:**

* es la aceleración gravitacional.
* es la masa total del sistema.
* son las aceleraciones lineal y angular, respectivamente.
* es la fuerza aplicada por el agente al carro.

El objetivo del sistema es mantener el péndulo en posición vertical mientras el carro permanece lo más cercano posible al centro del riel.

## Sistema de Recompensas

La función de recompensa está diseñada para incentivar la estabilidad tanto del péndulo como del carro. La expresión utilizada es la siguiente:

Donde:

* es el ángulo del péndulo.
* es la posición del carro.
* es el valor central deseado para .

Esta función penaliza tanto el ángulo (favoreciendo su estabilización) como la posición , para mantener el carro centrado.

## Algoritmo Q-Learning

El algoritmo Q-Learning se utiliza para aprender la política óptima del agente. Este algoritmo actualiza una tabla que almacena los valores esperados de las acciones en cada estado, usando la ecuación de Bellman:

**Donde:**

* es la tasa de aprendizaje.
* es el factor de descuento.
* es la recompensa obtenida en el estado después de realizar la acción .
* es el valor máximo esperado para el siguiente estado.

## Estados y Discretización

El sistema tiene 5 variables de estado continuas que definen el comportamiento del entorno:

1. ángulo del péndulo.
2. velocidad angular del péndulo.
3. posición del carro.
4. velocidad del carro.
5. aceleración del carro.

Estas variables se discretizan para facilitar el aprendizaje, considerando rangos y divisiones específicas para cada una:

* se discretiza en 10 valores entre y .
* se discretiza en 10 valores.
* se discretiza en 10 valores entre los límites del riel.
* se discretiza en 10 valores.
* se discretiza en 5 valores según las posibles aceleraciones.

El número total de estados es el producto de las discretizaciones de cada variable:

## Acciones

El agente tiene tres acciones posibles que puede realizar:

* **0**: Mover el carro hacia la izquierda.
* **1**: Mantener el carro en su posición actual (acción "quieto").
* **2**: Mover el carro hacia la derecha.

Cada acción impacta directamente en las dinámicas de y , y se utiliza para explorar el espacio de soluciones posibles.

## Políticas

Para balancear la exploración y explotación, se utiliza una política *epsilon-greedy*:

Donde disminuye gradualmente a medida que el agente gana experiencia, permitiendo una mayor explotación de la política aprendida.

## Selección de Hiperparámetros

Los hiperparámetros fueron seleccionados basándose en pruebas previas y teorías estándar (Sutton & Barto, 2018):

* **Tasa de aprendizaje** : Permite que el agente actualice los valores de de manera gradual.
* **Factor de descuento** : Permite valorar las recompensas futuras.
* **Tasa de exploración inicial** : Fomenta la exploración del espacio de acciones.

## Parámetros del Entorno y Fuerza

Los parámetros del entorno y la fuerza aplicada son los siguientes:

* Masa del carro:
* Masa del péndulo:
* Longitud del péndulo:
* Fuerza máxima aplicada:

La fuerza es discreta y puede tomar valores dentro de con incrementos de 20 N.

## Exploración y Explotación

El balance entre exploración y explotación es fundamental para lograr un aprendizaje eficiente. En las primeras etapas, la exploración predomina, con alto, permitiendo al agente probar diversas acciones. A medida que avanza el proceso de aprendizaje, la explotación aumenta, favoreciendo las acciones con mayor valor esperado en la tabla .

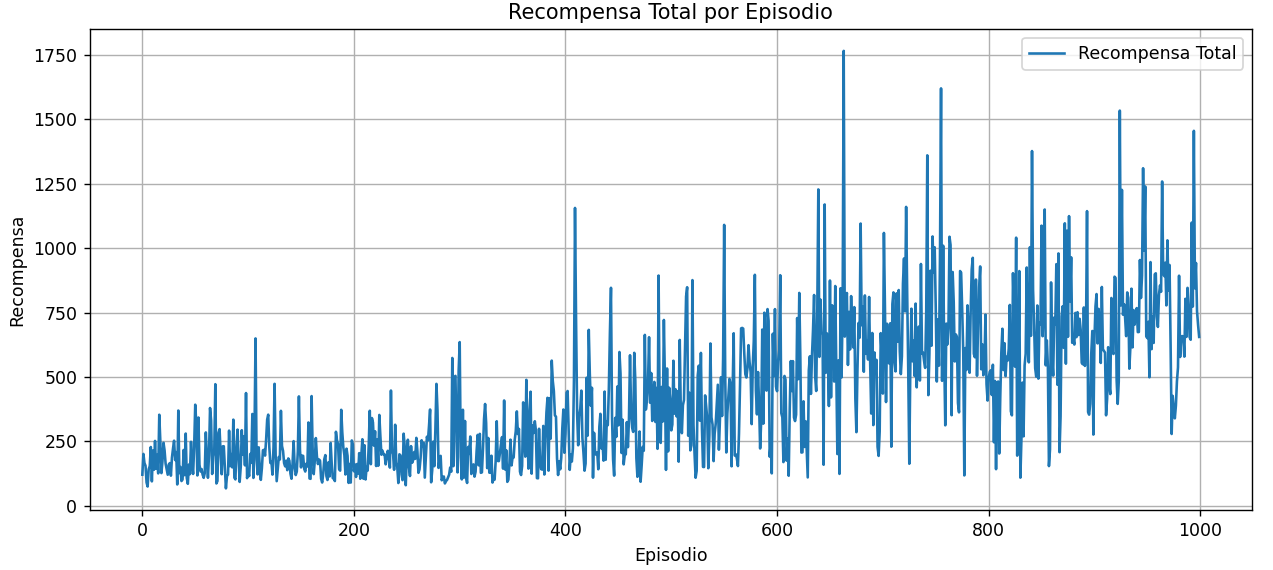
Capítulo 5: Validación y resultados

1. **Análisis de la Influencia de la Longitud del Péndulo y la Fuerza Máxima en el Rendimiento del Agente**

Se evaluaron tres configuraciones experimentales para determinar la influencia de la longitud del péndulo (LENGTH) y la fuerza máxima (MAX\_FORCE) en el rendimiento de un agente. Los resultados se presentan a continuación.

* Caso 1: **LENGTH = 100, MAX\_FORCE = 180.** Esta configuración produjo una recompensa promedio de 767.82 (Figura 1). Este resultado indica un rendimiento estable y eficiente, sugiriendo un balance óptimo entre la longitud del péndulo y la fuerza aplicada. La longitud permitió movimientos correctivos eficientes, mientras que la fuerza fue suficiente para mantener el equilibrio del péndulo.

Figura 1: Recompensa por Episodios (longitud = 100, fuerza\_max = 180)



Nota. *Gráfica de evolución de la recompensa Total por Episodio (1000).*

Describir el resultado de la imagen

* Caso 2: **LENGTH = 50, MAX\_FORCE = 360.** En este caso, se observó una disminución significativa en el rendimiento, con una recompensa promedio de 408.28 (Figura 2). La reducción en la longitud del péndulo incrementó su velocidad angular, haciendo el sistema más susceptible a errores. A pesar de la mayor fuerza máxima aplicada, no se compensó adecuadamente la alta velocidad angular, resultando en una inestabilidad y menor capacidad para mantener el equilibrio.

Figura 2:  Recompensa por Episodios (longitud = 50, fuerza\_max = 360)

Gráfico

Descripción generada automáticamente

Nota: *Gráfica de evolución de la recompensa total por Episodio (1000) .*

Descripción de la figura

* Caso 3: **LENGTH = 150, MAX\_FORCE = 90.** Con un péndulo más largo y una fuerza máxima reducida, la recompensa promedio fue de 443.76 (Figura 3). Si bien la mayor longitud facilitó el control al disminuir la velocidad angular, la limitación en la fuerza máxima dificultó la ejecución de correcciones rápidas y precisas, llevando a un rendimiento subóptimo.

Figura 3: Recompensa por Episodios (longitud = 150, fuerza\_max = 90)

Gráfico

Descripción generada automáticamente

Nota: *Gráfica de evolución de la recompensa total por Episodio (1000).*

Los resultados obtenidos indican la existencia de una relación crucial entre la longitud del péndulo y la fuerza máxima para optimizar el rendimiento del agente. Una longitud mayor proporciona mayor tolerancia a los errores debido a la menor velocidad angular; sin embargo, requiere una fuerza suficiente para realizar ajustes rápidos. Por el contrario, un péndulo corto es más difícil de controlar debido a su alta velocidad angular, y tanto el exceso como la insuficiencia de fuerza pueden generar inestabilidad. Por lo tanto, la selección cuidadosa de ambos parámetros es fundamental para lograr un desempeño óptimo y estable.

1. **Influencia de la Estructura de la Función de Recompensa en el Desempeño del Agente**

Se investigó el impacto de diferentes estructuras de la función de recompensa en el desempeño de un agente, utilizando tres configuraciones que variaban la penalización por el movimiento del carro. El entrenamiento se realizó durante 1000 episodios en cada configuración.

* Caso 1: **Penalización Estándar (k = 0.01).** La función de recompensa se definió como: . Este enfoque produjo una recompensa promedio de 607.16 (Figura 4), representando un balance adecuado entre la estabilidad del péndulo y el movimiento del carro.

Figura 4: Recompensa Total por Episodios

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio: 607.16*

* Caso 2: **Penalización Alta (k = 0.05).** Se incrementó la penalización por el movimiento del carro utilizando la función: . La recompensa promedio resultante fue de 598.27 (Figura 5), indicando una ligera disminución en el desempeño en comparación con el caso estándar. Este resultado sugiere que una penalización excesiva restringió innecesariamente el movimiento del carro, afectando negativamente el rendimiento general.

Figura5: Recompensa Total por Episodios

Gráfico, Gráfico de barras

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio:598.27*

* Caso 3: **Penalización Baja (k = 0.001).** Con una penalización reducida , la recompensa promedio obtenida fue de 357.60 (Figura 6). Este desempeño significativamente inferior al de los casos anteriores se atribuye a movimientos descontrolados del carro, comprometiendo la estabilidad del péndulo.

Figura 6: Recompensa Total por Episodios

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio: 357.60*

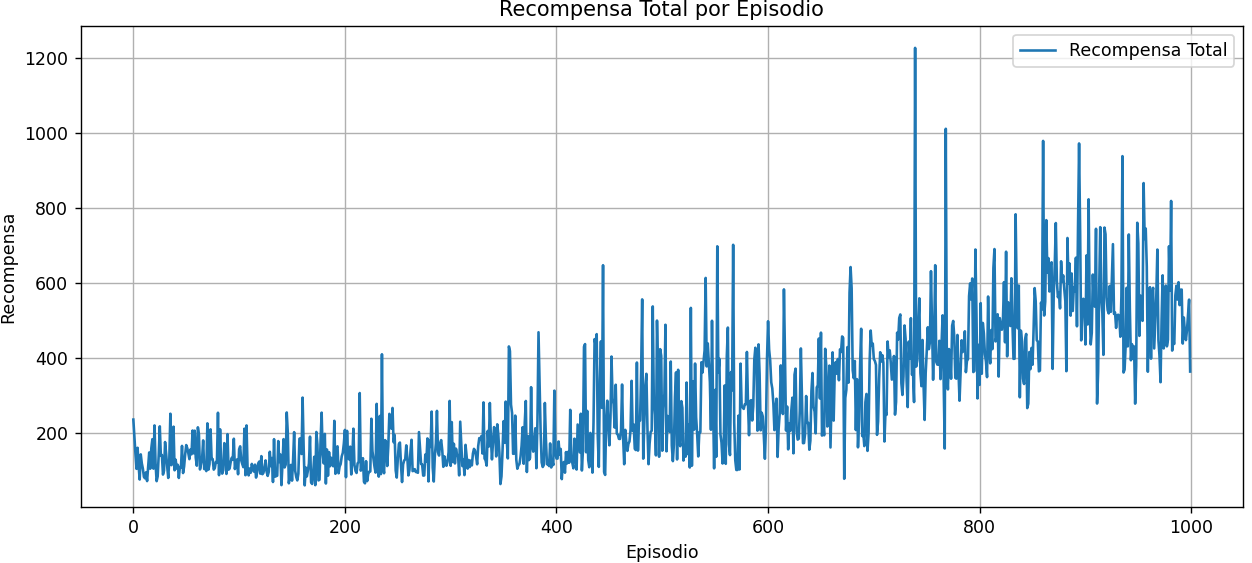
Los resultados muestran una clara dependencia del desempeño del agente en la magnitud de la penalización por el movimiento del carro. La configuración estándar (k = 0.01) ofreció el mejor rendimiento al lograr un equilibrio entre la estabilidad del péndulo y la libertad de movimiento del carro. Aumentar la penalización (k = 0.05) resultó en una ligera disminución del rendimiento, mientras que reducirla (k = 0.001) provocó un desempeño significativamente inferior. Estos hallazgos destacan la importancia de un ajuste preciso de la función de recompensa para optimizar el aprendizaje del agente y lograr un rendimiento satisfactorio.

1. **Tiempo de Convergencia del Agente y Efecto de los Hiperparámetros**

Se analizaron tres configuraciones de hiperparámetros para evaluar su impacto en el rendimiento de un agente de aprendizaje por refuerzo, enfocado en la estabilización de un sistema. Se compararon diferentes tasas de aprendizaje (learning\_rate, ), factores de descuento (gamma, ) y tasas de decaimiento de exploración (epsilon\_decay, ).

* Caso 1: **Configuración Estándar (, , ).** Esta configuración representó un balance entre velocidad de aprendizaje y estabilidad. Se obtuvo una recompensa promedio de 592.95 (Figura 7).

Figura 7: Recompensa Total por Episodios



Nota: *Episodios totales: 1000 y Recompensa promedio: 592.95*

* Caso 2: **Aprendizaje Rápido, Exploración Intensa (, , ).** Esta configuración priorizó la experiencia reciente (α alto) y las recompensas inmediatas (γ bajo). Se observó una recompensa promedio significativamente menor de 227.63 (Figura 8), indicando que, a pesar de la convergencia potencialmente más rápida, la mayor impulsividad del agente afectó negativamente la estabilidad del sistema.

Figura 8: Recompensa Total por Episodios

Gráfico

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio: 227.63*

* Caso 3: **Aprendizaje Lento, Exploración Reducida (, , ).** En este caso, se favoreció la estabilidad y un aprendizaje más gradual. La recompensa promedio obtenida fue de 230.67 (Figura 9), similar al caso de aprendizaje rápido. El bajo rendimiento se atribuye a la lentitud en la adaptación de las políticas del agente, priorizando la estabilidad a expensas de la eficiencia.

Figura 9: Recompensa Total por Episodios

Gráfico

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio: 230.67*

Los resultados indican que la elección de los hiperparámetros es crucial para el rendimiento del agente. La configuración estándar demostró un óptimo balance, ofreciendo el mejor rendimiento. Los casos de aprendizaje rápido y lento, aunque con objetivos contrastantes, resultaron en un desempeño significativamente inferior, destacando la importancia de una cuidadosa selección y ajuste de los parámetros α, γ y ε para lograr un equilibrio adecuado entre la velocidad de aprendizaje y la estabilidad del sistema.

1. **Respuesta del Agente a Estados Iniciales Desfavorables**

Se evaluó la capacidad del agente para recuperarse de estados iniciales desfavorables, utilizando tres configuraciones que simulaban diferentes situaciones desafiantes. En cada caso, se realizaron 1000 episodios.

* Caso 1: **Ángulo Extremo Cercano al Límite de Caída.** Se inicializó el sistema con un ángulo aleatorio entre y radianes (ángulo muy inclinado) y el carro en la posición central (). La recompensa promedio obtenida fue de 57.61 (Figura 10), indicando una dificultad significativa del agente para estabilizar el péndulo desde esta condición inicial. Se observaron fuertes oscilaciones, reflejando una baja capacidad de recuperación.

Figura 10: Recompensa Total por Episodios

Gráfico, Gráfico de barras

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa  promedio: 57.61*

* Caso 2: **Carro en Posición Extrema Izquierda.** Se inició con un ángulo θ cercano a la vertical (aleatorio entre -0.05 y 0.05 radianes) y el carro en una posición aleatoria entre 0 y 50 unidades de ancho (cerca del borde izquierdo). La recompensa promedio en este caso fue de 377.47 (Figura 11), mostrando una mejora considerable en comparación con el caso 1. La tendencia al alza sugiere una mayor capacidad del agente para recuperar el equilibrio desde esta posición.

Figura 11: Recompensa Total por Episodios

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio: 377.47*

* Caso 3: **Carro en Posición Extrema Derecha con Ángulo Inclinado.** Se simuló un estado inicial con un ángulo θ aleatorio entre y radianes (ángulo inclinado) y el carro en una posición aleatoria entre WIDTH - 50 y WIDTH (cerca del borde derecho). La recompensa promedio obtenida fue extremadamente baja, de 3.79 (Figura 12), indicando un desempeño deficiente del agente en esta situación. La tendencia cercana a cero sugiere una incapacidad para estabilizar el sistema.

Figura 12: Recompensa Total por Episodios

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

Nota: *Episodios totales: 1000 y Recompensa promedio: 37.9*

Los resultados muestran que el agente presenta dificultades significativas para manejar estados iniciales desfavorables, particularmente cuando se combinan ángulos de inclinación extremos con posiciones extremas del carro. El espacio disponible para maniobras parece ser un factor determinante en la capacidad del agente para estabilizar el sistema. El caso 2, con un ángulo inicial cercano a la vertical y el carro desplazado a la izquierda, demostró un desempeño considerablemente mejor, mientras que las situaciones que involucran ángulos pronunciados y posiciones extremas del carro resultaron en un rendimiento muy pobre. Esto indica áreas de mejora en el algoritmo de aprendizaje o en la función de recompensa para fortalecer la capacidad de recuperación del agente ante condiciones iniciales desafiantes.

1. **Evolución de la Recompensa Acumulada por Episodio Durante el Entrenamiento**

Se analizó la evolución de la recompensa acumulada por episodio durante el entrenamiento del agente, utilizando diferentes números de episodios. Se presentan los resultados para 500, 1000, 10000, 20000 y 40000 episodios.

* Caso 1: 500 Episodios. La Figura 13 muestra la evolución de la recompensa acumulada durante los primeros 500 episodios de entrenamiento.

Gráfico

Descripción generada automáticamente

Nota: *Evolución de la gráfica durante 500 episodios*

* Caso 2: 1000 Episodios. La Figura 14 ilustra la evolución de la recompensa durante 1000 episodios.

Figura 14: Recompensa Total por Episodios

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Nota: *Evolución de la gráfica durante 1000 episodios*

* Caso 3: 10000 Episodios. La Figura 15 presenta la recompensa acumulada para 10000 episodios.

Figura 15: Recompensa Total por Episodios

Gráfico

Descripción generada automáticamente

Nota: *Evolución de la gráfica durante 10000 episodios*

* Caso 4: 20000 Episodios. La Figura 16 muestra la evolución de la recompensa para 20000 episodios.

Figura 16: Recompensa Total por Episodios

Gráfico

Descripción generada automáticamente

Nota: *Evolución de la gráfica durante 20000 episodios.*

Descripción de la imagen

* Caso 5: 40000 Episodios. La Figura 17 presenta la recompensa acumulada después de 40000 episodios.

Figura 17: Recompensa Total por Episodios

Imagen que contiene Gráfico

Descripción generada automáticamente

Nota: *Evolución de la gráfica durante 40000 episodios.*

El análisis de las figuras muestra una clara relación entre el número de episodios de entrenamiento y la estabilidad de la recompensa acumulada. Con un número bajo de episodios (500 y 1000), el aprendizaje es inestable, con fuertes fluctuaciones. A medida que aumenta el número de episodios, la recompensa se estabiliza, convergiendo hacia un valor óptimo. A partir de 10000 episodios, se observa una convergencia casi completa, con fluctuaciones mínimas, indicando que el agente ha aprendido eficazmente a estabilizar el sistema. El incremento de episodios más allá de 10000 no produce una mejora significativa en el rendimiento, sugiriendo que se ha alcanzado un punto de saturación en el aprendizaje. Estos resultados resaltan la importancia de un entrenamiento adecuado para lograr un rendimiento óptimo en tareas de aprendizaje por refuerzo.

Capítulo 6: Conclusiones y Discusiones

El presente trabajo demuestra la viabilidad de aplicar el algoritmo de aprendizaje por refuerzo Q-learning para estabilizar un sistema de péndulo invertido modificado. A lo largo del desarrollo e implementación del modelo, se observaron patrones claros en la influencia de los parámetros físicos, la estructura de la función de recompensa y los hiperparámetros de aprendizaje en el rendimiento del agente.

**Conclusiones**

1. **Influencia de los Parámetros Físicos**

Los resultados confirman que existe una relación crítica entre la longitud del péndulo y la fuerza máxima aplicada. Un péndulo más largo favorece el control debido a su menor velocidad angular, pero requiere una fuerza suficiente para realizar ajustes correctivos. Por otro lado, péndulos más cortos incrementan la sensibilidad del sistema a errores debido a velocidades angulares altas, dificultando la estabilización. Este equilibrio se refleja en la superioridad del caso con LENGTH = 100 y MAX\_FORCE = 180, que obtuvo la recompensa promedio más alta (767.82), indicando un desempeño óptimo.

1. **Diseño de la Función de Recompensa**

La estructura de la función de recompensa influye significativamente en el desempeño del agente. Penalizaciones demasiado altas limitan la capacidad del agente para realizar movimientos correctivos efectivos, mientras que penalizaciones bajas permiten movimientos descontrolados, afectando negativamente la estabilidad del sistema. La configuración estándar (k=0.01) logró un balance adecuado, obteniendo la mejor recompensa promedio (607.16).

1. **Impacto de los Hiperparámetros**

La configuración estándar de hiperparámetros (α=0.1, γ=0.99, ϵ\_decay=0.995) presentó el mejor rendimiento, logrando un balance entre la velocidad de aprendizaje y la estabilidad. Por el contrario, configuraciones que priorizan un aprendizaje rápido o lento presentaron menor rendimiento debido a decisiones impulsivas o ajustes excesivamente conservadores, respectivamente.

1. **Resiliencia ante Estados Iniciales Desfavorables**

El agente mostró limitaciones para manejar estados iniciales extremos, como ángulos cercanos al límite de caída o posiciones extremas del carro. En estos casos, la recompensa promedio disminuyó significativamente, destacando la necesidad de mejorar la capacidad del sistema para adaptarse a condiciones adversas.

1. **Evolución del Entrenamiento**

Los gráficos de recompensa acumulada por episodio evidencian que el agente converge hacia una política estable en un número moderado de episodios bajo la configuración estándar. Sin embargo, tiempos de convergencia prolongados fueron observados en configuraciones más conservadoras.

**Discusiones**

El uso de Q-learning para el control de un péndulo invertido modificado representa una alternativa robusta y eficiente frente a los métodos de control convencionales. La capacidad del agente para aprender estrategias óptimas directamente de la interacción con el entorno elimina la necesidad de modelos matemáticos explícitos, lo que lo hace especialmente valioso en sistemas dinámicos no lineales y difíciles de modelar.

Sin embargo, este enfoque también presenta limitaciones inherentes. La discretización del espacio de estados, si bien simplifica el problema, puede introducir restricciones que afecten el desempeño del sistema, especialmente en condiciones iniciales extremas. Asimismo, la selección de hiperparámetros y la estructura de la función de recompensa requieren ajustes cuidadosos para optimizar el aprendizaje.

Finalmente, los hallazgos destacan la importancia de equilibrar los parámetros físicos del sistema y los componentes del modelo de aprendizaje para maximizar el rendimiento. Este equilibrio resulta crucial para garantizar un desempeño estable y eficiente, particularmente en sistemas inestables como el péndulo invertido.

El presente trabajo abre camino para futuras investigaciones, como la implementación de algoritmos más avanzados que puedan superar las limitaciones observadas y explorar aplicaciones más complejas en control dinámico.

Referencias

Ataka, N. A., Sandiwan, N. A., Tnunay, N. H., Utomo, N. D. R., & Cahyadi, N. A. I. (2023). Inverted Pendulum Control: A Comparative Study from Conventional Control to Reinforcement Learning. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi (JNTETI)*, *12*(3), 197-204.

<https://doi.org/10.22146/jnteti.v12i3.7065>

Dev, A., Chowdhury, K. R., & Schoen, M. P. (2024). Q-Learning Based Control for Swing-Up and Balancing of Inverted Pendulum. *2022 Intermountain Engineering, Technology And Computing (IETC)*, 209-214. <https://doi.org/10.1109/ietc61393.2024.10564347>

Israilov, S., Fu, L., Sánchez-Rodríguez, J., Fusco, F., Allibert, G., Raufaste, C., & Argentina, M. (2023). Reinforcement learning approach to control an inverted pendulum: A general framework for educational purposes. *PLoS ONE*, *18*(2), e0280071. <https://doi.org/10.1371/journal.pone.0280071>

Safeea, M., & Neto, P. (2023). A Q-learning approach to the continuous control problem of robot inverted pendulum balancing. *Intelligent Systems With Applications*, *21*, 200313.

<https://doi.org/10.1016/j.iswa.2023.200313>

Sharma, S. (2020). Modeling an Inverted Pendulum via Differential Equations and Reinforcement Learning Techniques. *Preprints*. <https://doi.org/10.20944/preprints202005.0181.v1>

Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press. <https://ieeexplore.ieee.org/book/6267343>