

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES
UNIVERSIDAD POLITÉCNICA DE MADRID

José Gutiérrez Abascal, 2. 28006 Madrid
Tel: 91 336 3060
info.industriales@upm.es

www.industriales.upm.es



POLITÉCNICA

INDUSTRIALES

05 TRABAJO FIN DE GRADO

Miguel Fernández Cortizas

TRABAJO FIN DE GRADO

**KIT DE DESARROLLO Y
VALIDACIÓN DE ALGORITMOS DE
CONTROL DE ACTITUD PARA
CUADRICÓPTEROS.**

SEPTIEMBRE 2019

Miguel Fernández Cortizas

**DIRECTOR DEL TRABAJO FIN DE GRADO:
Pascual Campoy Cervera**

UNIVERSIDAD POLITÉCNICA DE MADRID
ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INDUSTRIALES
GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES



POLITÉCNICA

KIT DE DESARROLLO Y VALIDACIÓN DE ALGORITMOS DE CONTROL DE ACTITUD PARA CUADRICÓPTEROS.

Miguel Fernández Cortizas

Tutor académico:

D. Pascual Campoy Cervera
D. Alejandro Rodríguez Ramos

Madrid - España
2019

*En memoria de mi padrino Juan,
seguiré trabajando hasta alcanzar las metas
que me hubiese gustado celebrar contigo.*

Agradecimientos

En primer lugar quiero darle las gracias a mis padres, por el cariño y el apoyo incondicional que me han brindado durante todos estos años y de los cuales, he aprendido a tener pasión por trabajar en lo que me gusta.

En segundo lugar, tengo de darle las gracias a Carmen, por estar a mi lado siempre que la necesito, apoyándome y ayudándome, siempre con una sonrisa.

Finalmente quiero darle las gracias Pascual Campoy por ofrecerme la oportunidad de realizar este trabajo, a Alejandro Rodríguez, por ayudarme, orientarme y enseñarme durante todo el transcurso de este trabajo, y al resto de compañeros del CVAR, gracias a los cuáles he tenido la oportunidad de aprender y trabajar en un ambiente inmejorable.

Resumen ejecutivo

Introducción

Los cuadricópteros son vehículos aéreos no tripulados (del inglés, *UAV*) que se emplean para una gran variedad de tareas en diversas áreas, donde cabe resaltar la inspección industrial (palas de aerogeneradores), operativos de búsqueda y rescate, cine, etc. Estas aeronaves son inherentemente inestables, por lo que es necesario que cuenten con un sistema que se encargue de mantener a la aeronave estable y facilitar su pilotaje. El desarrollo de los algoritmos de control que se ejecutan en estos sistemas compone un campo de interés científico notable, ya que su mejora u optimización puede derivar en un nivel mayor de autonomía para los UAVs, lo que permitiría una mejor penetración de esta tecnología en las diversas áreas anteriormente mencionadas.

Alcance

Los objetivos de este trabajo son: diseñar una plataforma real segura, en la que se pueda probar y evaluar el rendimiento de distintos algoritmos de control de cuadricópteros, así como diseñar y probar algoritmos de control novedosos empleando las últimas técnicas de aprendizaje por refuerzo y aprendizaje profundo. Esta plataforma puede ser usada tanto para la labor investigadora (a la hora de desarrollar nuevos algoritmos de forma segura), como para la labor docente (enseñar las técnicas de control clásicas en una plataforma real).

Para acometer este proceso, se ha desarrollado un autopiloto en el que poder ejecutar los algoritmos de control que se diseñen, a bordo de la aeronave y un cuadricóptero en el que realizar los experimentos.

Adicionalmente, se ha empleado un entorno de simulación en el que se han desarrollado y validado los distintos algoritmos de aprendizaje por refuerzo que se han propuesto. El código empleado se encuentra en el repositorio: <https://github.com/mifero97/gymfc>

Solución realizada

- **Diseño del autopiloto.** Un autopiloto es el sistema encargado de estabilizar y comandar a la aeronave a bajo nivel. Con el objetivo de poder implementar los algoritmos de control sin restricciones, se ha diseñado un autopiloto propio con toda la electrónica necesaria para poder controlar un cuadricóptero fig. 1.

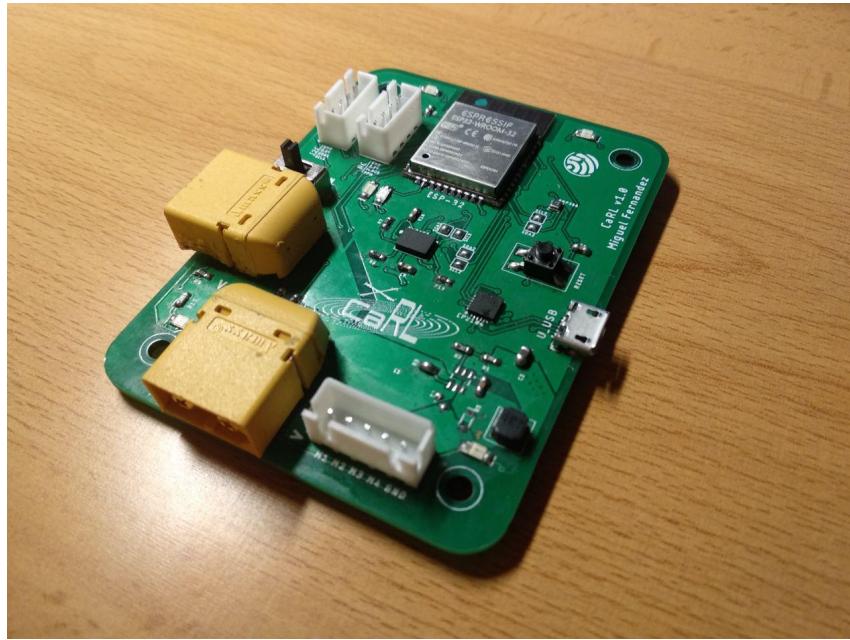


Figura 1: Autopiloto CaRL (*Cuadcopter with autopilot based on Reinforcement Learning*).

- **Diseño del cuadricóptero y el banco de pruebas.** Para poder evaluar el rendimiento de los algoritmos de control empleados, es necesario contar con una plataforma real, en la que sea seguro probar estos algoritmos. Es por esto que se ha diseñado y construido un cuadricóptero, en el que montar el autopiloto y un banco de pruebas con diversas configuraciones, dependiendo del experimento que se quiera realizar.

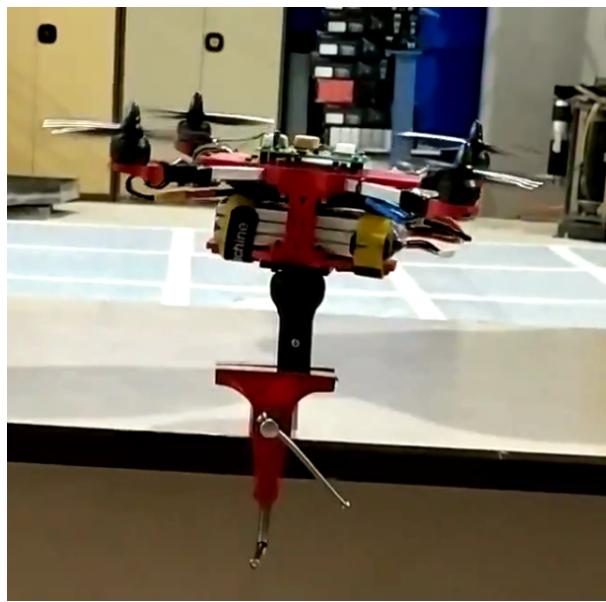


Figura 2: Cuadricóptero sobre el banco de pruebas

- **Desarrollo de algoritmos de control basados en aprendizaje por refuerzo.** Con el ánimo de encontrar métodos de control novedosos, que sean capaces de controlar a los cuadricópteros de forma más eficaz, se ha experimentado con los distintos algoritmos de estado del arte en el campo del aprendizaje por refuerzo.

Experimentación y resultados.

Se han realizado experimentos tanto en la plataforma real como en simulación. En la plataforma real se ha probado el funcionamiento del autopiloto, la aeronave y el banco de pruebas empleando algoritmos de control clásico, con los que se ha conseguido estabilizar la plataforma en diferentes configuraciones. Los vídeos de los experimentos reales se encuentran disponibles en <https://vimeo.com/359169626>.

En el entorno simulado, se ha comparado el rendimiento de los algoritmos de control empleando aprendizaje por refuerzo, con el algoritmo de control clásico PID (Proporcional-Integral-Derivativo). En estos experimentos se ha observado como alguno de los algoritmos empleados, como el TRPO, son capaces de conseguir respuestas con características dinámicas muy satisfactorias.

Conclusiones y trabajo futuro

Se ha conseguido desarrollar una plataforma que permite probar de forma segura distintos algoritmos de control en cuadricópteros, esta plataforma se podría redimensionar para poder ser empleada en institutos y universidades con el ánimo de facilitar el aprendizaje de la teoría de control mediante la interacción directa con un sistema tan atractivo, como el de un cuadricóptero.

Desde el punto de vista del algoritmo de control, se ha conseguido entrenar agentes capaces de estabilizar un cuadricóptero en simulación con resultados prometedores, eso abre el paso a intentar mejorar estos algoritmos con el propósito de conseguir controladores mas precisos y eficaces.

Palabras clave

UAV, cuadricóptero, control clásico, aprendizaje automático, inteligencia artificial, aprendizaje por refuerzo, redes neuronales.

Códigos UNESCO

120304 INTELIGENCIA ARTIFICIAL

120326 SIMULACIÓN

330104 AERONAVES

330412 DISPOSITIVOS DE CONTROL

330703 DISEÑO DE CIRCUITOS

Índice general

1	Introducción	10
1.1	Motivación	10
1.2	Solución propuesta	11
1.3	Objetivos	12
2	Estado del arte	13
2.1	Plataformas de control de cuadricópteros	13
2.1.1	Plataformas tipo gimbal	13
2.1.2	Plataformas con unión esférica	14
2.2	Aprendizaje por refuerzo	16
3	Fundamentos teóricos	17
3.1	Controlador PID	17
3.2	Redes neuronales artificiales	19
3.3	Aprendizaje por refuerzo	21
3.3.1	Algoritmos de <i>Q-learning</i>	23
3.3.2	Algoritmos de gradiente de política	25
4	Hardware	28
4.1	Cuadro	29
4.2	Motores y hélices	31
4.3	Variadores (ESC)	32
4.4	Baterías	33
4.5	Autopiloto	33
4.5.1	Fase de Potencia	34
4.5.2	El microcontrolador (ESP32)	35
4.5.3	Sensores	36
4.6	Banco de pruebas	36
5	Software	39
5.1	Software del Autopiloto	39
5.1.1	Estimación del Estado	39
5.1.2	Interfaz WiFi	39
5.1.3	Generacion de comandos <i>on-board</i>	40
5.2	Software de la estación	40
5.2.1	Entorno de simulación	40
5.2.2	Agente (generación de comandos)	42
5.2.3	Interfaz Estación-Autopiloto	42
5.3	Descripción del equipo	43

6 Metodología	44
6.1 Diseño del estado	44
6.2 Diseño de las acciones	44
6.3 Diseño de la función de recompensa y ajuste de hiperparámetros	45
7 Experimentos	47
7.1 Experimentos en simulación	47
7.2 Experimentos en real	48
8 Conclusiones y trabajo futuro	53
8.1 Conclusiones	53
8.2 Trabajo futuro	54
A Esquemáticos del autopiloto	55
B Presupuesto y Planificación	59
B.1 Presupuesto	59
B.2 Planificación	60
C Impacto social y medioambiental	62

Introducción

Un multirrotor es un vehículo aéreo no tripulado o UAV (por sus siglas en inglés *Unmanned Aerial Vehicles*) cuyos motores y hélices están orientadas de forma vertical. Estas aeronaves son mucho más maniobrables que una aeronave de ala fija, ya que, al poder mantenerse estables en una posición fija, pueden realizar trayectorias de forma más precisa en espacios reducidos. Comparado con un helicóptero, el cual también puede mantenerse estable en un punto fijo, los multirrotos cuentan con un mantenimiento mucho más sencillo debido a la ausencia de complejos mecanismos.

Generalmente los motores de estas aeronaves son eléctricos, ya que poseen una gran velocidad de reacción y son capaces de manejar una gran cantidad de energía con rendimientos muy elevados. La energía que requieren se proporciona a través de baterías, lo cual limita el tiempo de vuelo de estas aeronaves, cuya autonomía es significativamente inferior a la de las aeronaves de ala fija o a la de los helicópteros con motores de combustión.

Todo esto, unido a su reducido precio, ha permitido que se potencie el uso masivo de este tipo de aeronaves para labores de: agricultura de precisión, topografía, inspección industrial y rodajes cinematográficos, entre otros usos. Existen diversas topologías asociadas a los multirrotos, dependiendo del número de motores y la disposición de éstos. En este trabajo se ha desarrollado un cuadricóptero o cuadrirrotor (cuenta con cuatro motores) con disposición en X (Fig. 1.1).

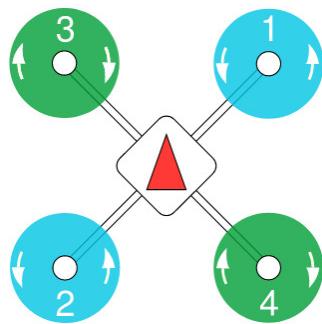


Figura 1.1: Esquema cuadricóptero en X

Se ha escogido un multirrotor de cuatro motores debido a que posee el mínimo número de motores que permiten un modelo sencillo de la dinámica de la aeronave.

1.1. Motivación

La popularidad de estas aeronaves en ámbitos como la inspección, la topografía o el ámbito cinematográfico requieren que la nave sea muy estable y que se pueda manejar con precisión. Para este propósito, todas estas aeronaves cuentan con un sistema electrónico

que se programa para que la nave sea estable y que facilite el pilotaje de la misma. A este sistema se le conoce como la controladora de vuelo o autopiloto.



Figura 1.2: Autopiloto comercial Pixhawk 4

Actualmente, los autopilotos emplean algoritmos de control clásicos, basados la gran mayoría en reguladores PID (Proporcional-Integral-Derivativo) para asegurar la estabilidad de la aeronave y permitir que se pilote de forma precisa. Debido a la peligrosidad de los drones (hélices girando a miles de revoluciones por minuto), es altamente complejo probar algoritmos de control novedosos en un cuadricóptero, sin tenerlo anclado a una estructura base, debido a que se trata de un sistema inestable de forma inherente en bucle abierto.

Las teoría clásica de control empleada para estabilizar un cuadricóptero requiere de un conocimiento preciso del modelo del sistema, junto con un fino ajuste de los parámetros del controlador. En este sentido, las últimas líneas de investigación apuntan a técnicas de aprendizaje, tanto para la calibración de los parámetros de un controlador, como para la realización de la funcionalidad completa de control. En el presente proyecto, se estudiarán algunas de las técnicas más novedosas en este campo.

1.2. Solución propuesta

Con el objetivo de poder desarrollar nuevos algoritmos para el control de cuadricópteros, se ha desarrollado una plataforma tanto hardware como software, que permite diseñar y probar estos algoritmos de forma segura. Esta plataforma esta constituida de:

- **Entorno de simulación**, sobre el que se puedan probar los algoritmos de control en una aeronave simulada.
 - **Cuadricóptero con autopiloto de diseño propio**, será la aeronave donde se probarán los algoritmos diseñados. Al contar con una controladora de vuelo de diseño propio se pueden implementar los distintos algoritmos de control a bordo.
 - **Interfaz con el autopiloto**, el cual comunica a la aeronave con un ordenador, capaz de enviar comandos a la aeronave y recibir el estado de la misma. Esto permite la implementación de algoritmos *en tierra*, en los cuales la computación del algoritmo se realiza en un ordenador en vez de en el microcontrolador del autopiloto.
 - **Banco de pruebas**, con distintas configuraciones en función del experimento que se desee realizar. Este banco permitirá probar los algoritmos de control de forma segura.

Para la validación de la solución propuesta, se ha probado el rendimiento de distintos algoritmos de control basados en técnicas de aprendizaje por refuerzo para poder compararlos con los algoritmos de control basados en el método de PID.

1.3. Objetivos

Para poder llevar a cabo esta plataforma y poder implementar algún algoritmo basado en aprendizaje por refuerzo es necesario desgranar los objetivos principales en tareas de alcance más reducido:

- **Entorno de simulación:**

- Estudio del estado del arte acerca de entornos de simulación para el diseño de algoritmos de control para cuadricópteros.
- Estudio del estado del arte acerca de las técnicas de aprendizaje por refuerzo y el control de cuadricópteros.
- Adaptación del entorno de simulación a la plataforma escogida e integración con ROS.
- Implementación de algoritmos clásicos en el entorno de simulación.
- Diseño de nuevos algoritmos de control basados en aprendizaje por refuerzo y entrenamiento de los mismos.
- Comparativa de los algoritmos en simulación.

- **Cuadrirrotor con autopiloto:**

- Estudio del estado del arte acerca de los autopilotos comerciales.
- Diseño y construcción completa del autopiloto: componentes, placa de circuito impreso y soldadura de componentes.
- Programación del *firmware* del autopiloto.
- Diseño CAD de los componentes mecánicos del cuadrirrotor e impresión 3D de los mismos.
- Montaje y ensamblaje de todos los componentes de la aeronave.

- **Interfaz con el autopiloto:**

- Diseño del protocolo de comunicación WiFi
- Integración del protocolo con ROS y con el autopiloto.

- **Banco de pruebas:**

- Diseño CAD de las piezas de los distintos bancos de pruebas e impresión 3D de las mismas.

- **Experimentación de los algoritmos en el mundo real:**

- Implementación a bordo de los algoritmos clásicos.
- Implementación externa de los algoritmos clásicos.
- Implementación externa de los algoritmos basados en técnicas de aprendizaje por refuerzo.

Estado del arte

El desarrollo del trabajo realizado se puede subdividir en dos campos: la construcción de la plataforma de sujeción del cuadricóptero y la experimentación con nuevos algoritmos de control empleando técnicas de aprendizaje por refuerzo. Es por esto que, con el ánimo de contextualizar este trabajo dentro de estos dos campos, se ha tratado el estado del arte de forma separada.

2.1. Plataformas de control de cuadricópteros

Debido a la creciente popularidad de estas aeronaves, muchas líneas de investigación han trabajado en probar y desarrollar distintos algoritmos de control, para comparar el rendimiento entre ellos. Para poder testear estos algoritmos es muy conveniente contar con una plataforma de control, que te permita medir el rendimiento de estos algoritmos de control de forma precisa y segura.

Principalmente, existen 2 tipos de plataformas para este propósito: Plataformas tipo gimbal y estructuras con unión esférica. A continuación se tratará mas detalladamente las características de estas estructuras:

2.1.1. Plataformas tipo gimbal

Estas plataformas están formados por 3 anillos unidos entre ellos dos a dos, de tal forma que el anillo interior gira respecto al anillo que lo rodea y éste a su vez gira sobre el anillo exterior, como se puede observar en la figura 2.1

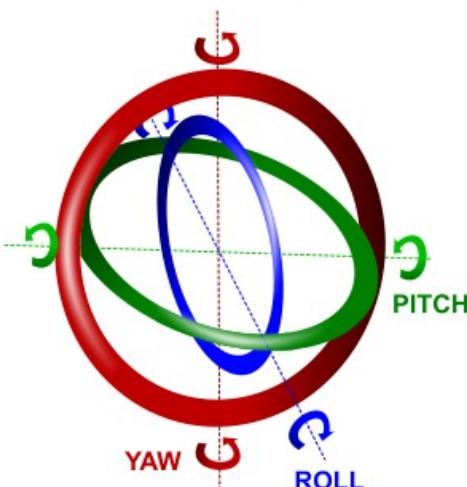


Figura 2.1: Esquema estructura gimbal

Estas plataformas permiten que la aeronave gire entorno a su centro de gravedad, por lo que se reduce la influencia del peso de la aeronave en el control. Además es posible medir los ángulos de euler de la aeronave de forma directa empleando encoders en los ejes de rotación.

Hay empresas que están empezando a comercializar estas plataformas, tanto orientadas para la labor investigadora en el campo de los cuadricópteros, como para la labor docente que se puede llevar a cabo empleando estas plataformas con ánimo didáctico, para el aprendizaje de algoritmos de control.

En cuanto a las plataformas orientadas a la investigación, se encuentra la plataforma FFT gyro , desarrollada por Eureka Dynamics.



Figura 2.2: Plataforma FFT Gyro de Eureka Dynamics

Esta plataforma cuenta con encoders en cada eje de rotación para poder monitorizar el estado de la aeronave con gran precisión y esta construida en fibra de carbono para maximizar la rigidez de la estructura minimizando el peso.

2.1.2. Plataformas con unión esférica

Estas plataformas se unen a la aeronave a través de una unión esférica, la cual permite que el cuadricóptero pueda girar en tres ángulos, con algunas limitaciones. Por ejemplo, aunque en guinada la aeronave pueda girar libremente, en alabeo y cabeceo este movimiento se ve restringido por las limitaciones mecánicas de la unión, véase fig. 2.4. Esto es una gran diferencia con respecto a las plataformas de tipo gimbal, en las que este suceso no ocurre, dado que en sus articulaciones no existen restricciones en los ángulos de giro.

A pesar de esto, estas plataformas condensan todos los giros en un único punto, por lo que el tamaño de la estructura requerida para anclar un cuadricóptero, es significativamente menor que si se emplea una estructura de tipo gimbal.

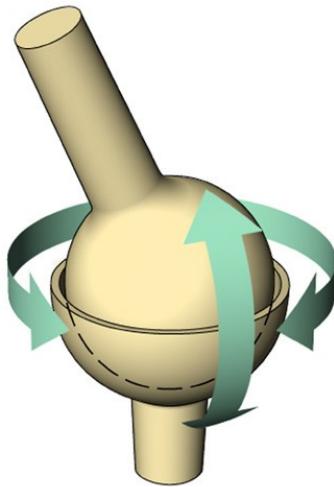


Figura 2.3: Esquema de una union esférica

En cuanto a plataformas comerciales de este tipo, cabe destacar las producidas por la empresa Quanser, la cuál ha desarrollado una plataforma con fines educativos, que emplea esta unión. Esta plataforma cuenta, al igual que la FFT Gyro con encoders, con la finalidad de conocer con precisión el estado de la aeronave.



Figura 2.4: Plataforma 3 DOF Hover de la empresa Quanser

Entre las dos posibilidades, se ha optado por realizar una plataforma con una unión esférica debido a la mayor simplicidad y menor tamaño de este tipo de plataformas frente a su alternativa.

2.2. Aprendizaje por refuerzo

El aprendizaje por refuerzo o *reinforcement learning* es una rama del aprendizaje automático en la que un agente aprende a actuar a medida que va interactuando con su entorno. Es decir, el agente comienza realizando acciones aleatorias y va aprendiendo por ensayo-error. Para que el agente tenga noción de que acciones están “bien” y cuales no, el entorno proporciona una recompensa al agente en función de su comportamiento. Esta rama del aprendizaje automático se inspira en la psicología conductista.

Debido a esta capacidad de aprender de forma autónoma, sin necesidad de un conjunto previo de datos etiquetados, sino extrayendo información únicamente de su interacción con el entorno, estos algoritmos son muy empleados en el campo de la robótica para llevar a cabo tareas complicadas, tales como, aprender a andar o a manipular un cubo con destreza [1].

Aunque existen una gran cantidad de ejemplos de uso de estas técnicas de aprendizaje en el ámbito de la robótica, en este trabajo se ha centrado en el empleo de estas técnicas al control de UAVs.

Los primeros experimentos en la utilización de algoritmos de control para UAVs, entrenados con aprendizaje con refuerzo, se llevaron a cabo con helicópteros. En 2004, HJ Kim et al. [2] emplearon algoritmos de *reinforcement learning* para estabilizar (*hover*) un helicóptero y conseguir realizar maniobras acrobáticas, para ello desarrollaron el algoritmo Pegasus [3]. Años después, en 2006 Andrew Y. et al [4] continuaron la investigación, en esta ocasión emplearon algoritmos que aprendían a realizar maniobras acrobáticas complejas, como mantener invertido al helicóptero. Para ello, realizaron un modelo estocástico no lineal de la dinámica del helicóptero, para, posteriormente, emplear ese modelo en simulación con el ánimo de generar un controlador capaz de estabilizar al helicóptero invertido.

En 2010 Travis Dierks et al. [5] desarrollaron un controlador no lineal, basado en redes neuronales, para estabilizar un cuadricóptero y seguir trayectorias. Para obtener un modelo completo de la aeronave, que fuera capaz también de tener en cuenta otros parámetros externos a la aeronave, como el coeficiente aerodinámico, el aprendizaje de la red neuronal se realizaba a tiempo real, mientras el cuadricóptero volaba.

Unos años después, en 2017 Jemin Hwangbo et al. [6] desarrollaron un método para controlar un cuadricóptero con una red neuronal usando técnicas de *reinforcement learning*. Desarrollaron un algoritmo, basado en la optimización determinista de la política y empleando descenso de gradientes natural para optimizar la misma. Consiguieron que el cuadricóptero fuera capaz de estabilizarse aún cuando partía de condiciones adversas, como ser lanzado casi boca abajo contra el suelo. Sin embargo, para conseguirlo, necesitaban frecuencias de actualización muy elevadas, del orden de los 100Khz, dos órdenes de magnitud mayor que un algoritmo típico.

En 2018 William Koch et al. [7] desarrollaron un entorno de simulación, GYMFC, para el desarrollo de nuevos algoritmos de control basados en redes neuronales. En su artículo compararon el rendimiento, en simulación, de distintos algoritmos de aprendizaje por refuerzo a la hora de cerrar el bucle de control en velocidad de un cuadricóptero. Posteriormente, a comienzos de 2019, desarrollaron Neuroflight [8], un *firmware* para autopilotos *open-source*. Este *firmware*, a diferencia del de los autopilotos convencionales emplea bucles de control desarrollados con técnicas de aprendizaje por refuerzo, con los que consiguen controlar un cuadricóptero real.

Fundamentos teóricos

3.1. Controlador PID

Un regulador PID es un controlador lineal realimentado. Matemáticamente se expresa como

$$u(t) = \underbrace{K_p e(t)}_{\text{P}} + \underbrace{K_i \int_0^t e(\tau) d\tau}_{\text{I}} + \underbrace{K_d \frac{de(t)}{dt}}_{\text{D}} \quad (3.1)$$

donde $e(t)$ es el error entre el valor de consigna y el valor actual de la variable a medir, $u(t)$ representa la variable de control y K_p, K_i, K_d son parámetros ajustables de los cuales depende la dinámica y la estabilidad del sistema. Como se puede observar, el regulador consta de tres partes: la parte proporcional (P) tiene en cuenta el error actual, la parte integral (I) tiene en cuenta la acumulación histórica de los errores y la parte derivativa (D) tiene en cuenta el “futuro” del error.

En este trabajo se han empleado controladores PID en dos tipos de arquitecturas diferentes. La primera arquitectura consta de un bucle de control individual, donde el error de entrada de PID es calculado directamente con la diferencia de la variable a controlar y la medida. La fig. 3.1 presenta el diagrama de bloques de esta arquitectura. Esta estrategia es la más básica de las estructuras PID.

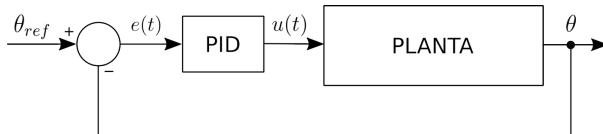


Figura 3.1: Bucle de control realimentado con regulador PID

La otra variante emplea una arquitectura de control en cascada, la cuál está integrada mediante dos bucles cerrados de control, uno interno y uno externo. El bucle externo emplea un regulador P para generar la señal de referencia utilizada por el bucle interno de control. El bucle interno es diseñado con un regulador PID para controlar la magnitud de una variable interna. En el caso particular de este trabajo, el bucle externo intenta alcanzar una referencia de posición angular θ_{ref} , para ello el regulador P es ajustado para generar una referencia de velocidad angular $\dot{\theta}_{ref}$ a un controlador de velocidad con un regulador PID, véase fig. 3.2.

Control de la aeronave

Para el control de la aeronave es necesario implementar 3 sistemas de control independientes, uno para cada ángulo. Cada bucle de control envía una señal $u(t)$ a la planta, en

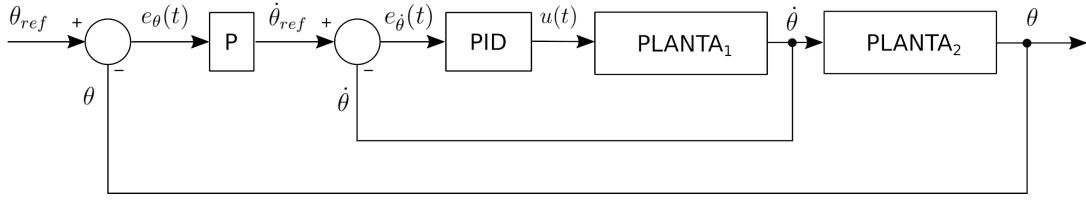


Figura 3.2: Bucle de control en cascada

este caso, a la aeronave. Para traducir estas señales a los comandos que se le envían a los motores, es necesario realizar una transformación.

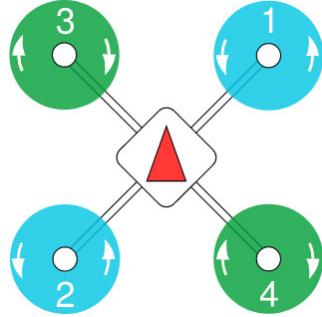


Figura 3.3: Esquema cuadrirrotor en X

Intuitivamente, si se desea que el cuadricóptero se incline hacia delante, la velocidad de los 2 motores traseros deberá aumentar, mientras que la velocidad de los motores delanteros deberá disminuirse. Si consideramos que inclinarse hacia delante implica que la nave tenga un ángulo $\theta > 0$ y que los motores están dispuestos según la figura 3.3, entonces

$$\begin{aligned} w_1 &= -u_\theta(t) \\ w_2 &= +u_\theta(t) \\ w_3 &= -u_\theta(t) \\ w_4 &= +u_\theta(t) \end{aligned}$$

Si lo expresamos matricialmente:

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \begin{bmatrix} -1 \\ +1 \\ -1 \\ +1 \end{bmatrix} u_\theta(t) \quad (3.2)$$

siendo w_i la velocidad del motor i y $u_\theta(t)$ la salida del controlador de *pitch*. Si se cierran los 3 sistemas de control de forma simultánea y se suman las contribuciones de cada bucle sobre las acciones de control resulta en

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \end{bmatrix} = \underbrace{\begin{bmatrix} -1 & -1 & +1 \\ +1 & +1 & +1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \end{bmatrix}}_{\text{Matriz de transformación}} \begin{bmatrix} u_\varphi(t) \\ u_\theta(t) \\ u_\psi(t) \end{bmatrix} \quad (3.3)$$

Esta matriz de transformación relaciona las salidas de los controladores con la acción de control requeridas por los actuadores, i.e., velocidad de rotación. Modificando estos valores se puede aumentar la influencia de un bucle con respecto a otro.

3.2. Redes neuronales artificiales

Una red neuronal artificial (ANN) esta compuesta por un conjunto de nodos o perceptrones interconectados entre sí. Estos perceptrones se agrupan en capas “ocultas”, se les atribuye este nombre debido a que todos los nodos de una capa se interconectan con todos los nodos de la capa anterior, por lo que después del aprendizaje de la red no se sabe cuales son los perceptrones de la capa anterior que influyen en un nodo.

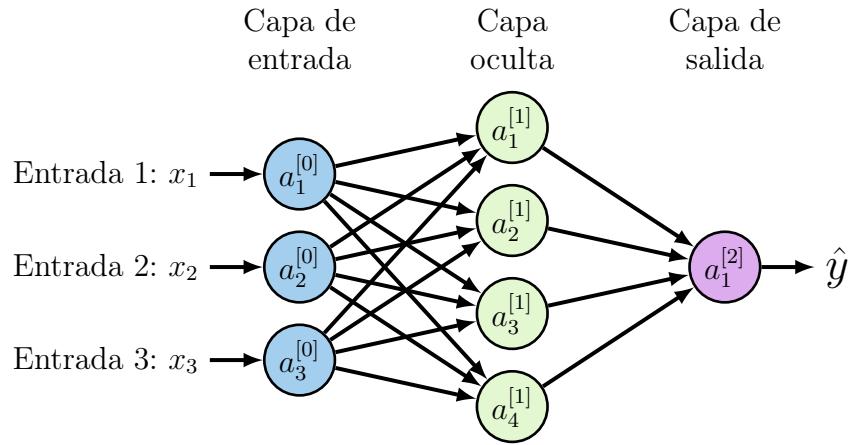


Figura 3.4: Esquema de una red neuronal artificial

Cada perceptrón es la unidad mínima de computación de una ANN, estas unidades se dividen en dos partes, una parte lineal y una parte de activación o no lineal. En la parte lineal o parte “Z” se computa una regresión lineal de las salidas de los nodos anteriores.

$$z_i^{[l]} = \sum_{j=0}^{n^{[l-1]}} w_{ij} \cdot a_j^{[l-1]} + b_i \quad i = 0, \dots, n^{[l]} \quad (3.4)$$

$$a_i^{[l]} = g(z_i^{[l]}) \quad i = 0, \dots, n^{[l]} \quad (3.5)$$

El superíndice $[l]$ hace referencia a la capa en la que se encuentra el elemento. Siendo $n^{[l]}$ el número de nodos de la capa l -ésima.

A los coeficientes w_{ij} se les denomina los pesos del perceptrón y b_i es el término independiente de la regresión.

La función $g(z)$ es la función de activación del nodo. Estas funciones proporcionan no linealidad a la red neuronal, permitiendo a estas la capacidad de generar modelos con grandes no linealidades. Las funciones de activación más frecuentes en la literatura son:

- Función sigmoide:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad \sigma(z) : \mathbb{R} \rightarrow [0, 1] \quad (3.6)$$

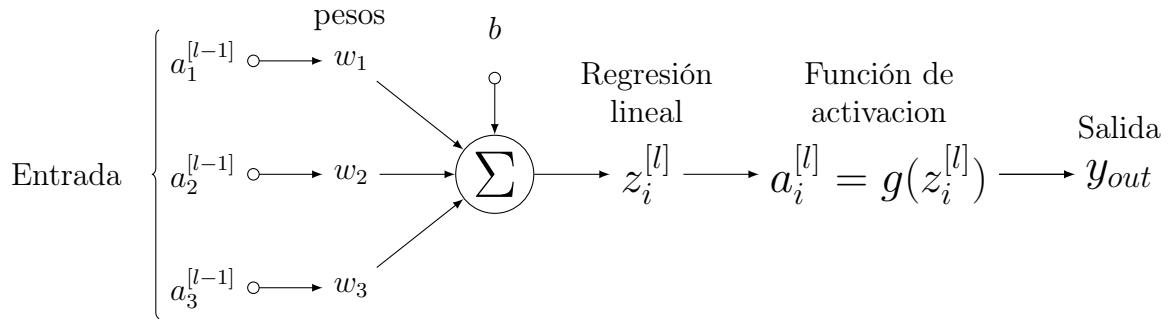
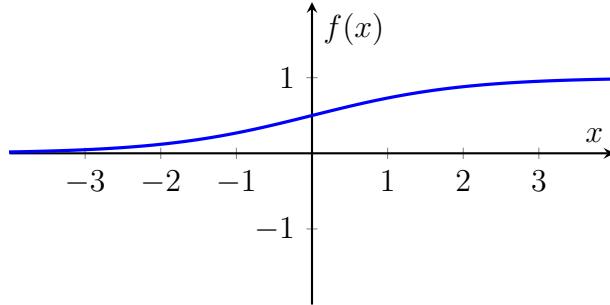
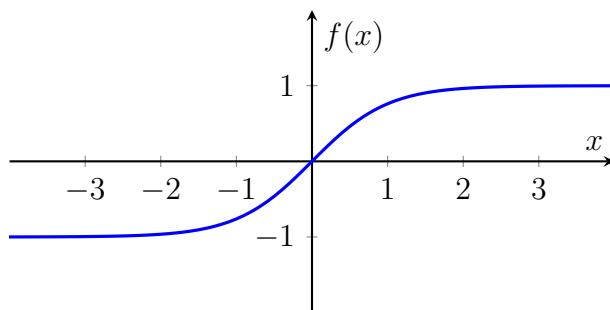


Figura 3.5: Esquema de un perceptrón



- Tangente hiperbólica:

$$\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \quad \tanh(z) : \mathbb{R} \rightarrow [-1, 1] \quad (3.7)$$



- ReLu (del inglés *Rectified Linear Unit*):

$$g(z) = \max(0, z) \quad g(z) : \mathbb{R} \rightarrow [0, +\infty] \quad (3.8)$$

Para conseguir que la red neuronal realice predicciones precisas es necesario ajustar los pesos de la red, a este proceso es al que se denomina entrenamiento o aprendizaje. En el paradigma del aprendizaje supervisado este entrenamiento se realiza sometiendo a la red a ejemplos cuya salida es conocida. El objetivo de la red es minimizar el error de

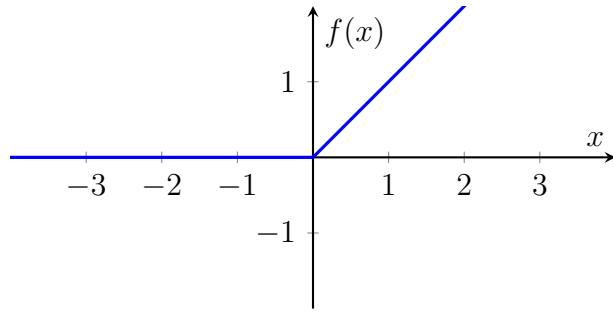


Figura 3.8: Función ReLu

la esperanza de la estimación con respecto a la salida real del ejemplo. Formalmente, se define una función de coste \mathcal{J} la cuál se quiere minimizar. Por ejemplo, una función de coste típica para problemas de clasificación binaria es la *binary cross-entropy*:

$$\mathcal{L}(\hat{y}, y) = -(y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \quad (3.9)$$

$$\mathcal{J}(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) \quad (3.10)$$

donde \hat{y} denota la estimación de la salida realizada por parte de la red, y la salida conocida y m el número de ejemplos.

Para minimizar esta función de coste, que depende de los pesos de la red, existen distintos métodos, uno de los más usados es el método del descenso de gradiente. Este método emplea la “propagación hacia atrás” (del inglés *back propagation*) de la red neuronal, esto consiste en obtener las derivadas parciales de la función de coste con respecto a los pesos de cada nodo y actualizar estos pesos en la dirección opuesta al máximo gradiente.

$$w := w - \alpha \frac{\partial \mathcal{J}(w, b)}{\partial w} \quad (3.11)$$

$$b := b - \alpha \frac{\partial \mathcal{J}(w, b)}{\partial b} \quad (3.12)$$

donde α denota la tasa de aprendizaje, es decir, lo rápido que varían estos pesos.

3.3. Aprendizaje por refuerzo

El aprendizaje por refuerzo o *Reinforcement learning* [9] es un área del aprendizaje automático o *Machine Learning* en el que un agente interactúa con un entorno buscando la mejor acción a realizar en función de su estado actual, de manera que maximice las recompensas acumuladas en el tiempo.

Se diferencia de otras técnicas de aprendizaje automático en su enfoque orientado a la interacción directa con el entorno, sin basarse en un modelo completo del entorno o en un conjunto de ejemplos supervisados.

Elementos del aprendizaje por refuerzo

Además del agente y el entorno se pueden identificar tres elementos principales más en un sistema de aprendizaje con refuerzo:

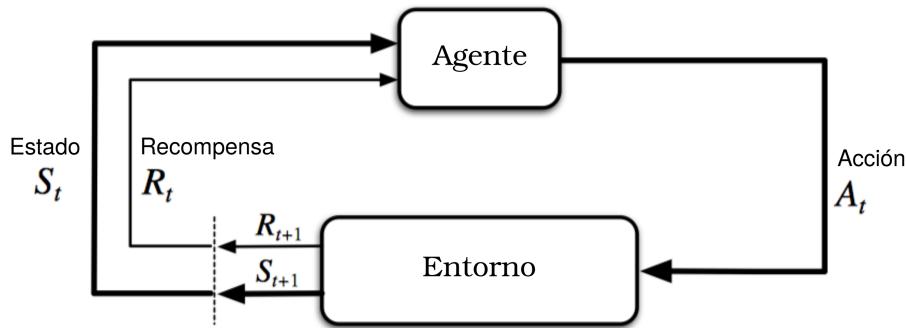


Figura 3.9: Diagrama canónico del bucle de interacción entorno-agente

- **Política (π)** (Proveniente del término anglosajón *policy*, el cual es el término empleado en el estado del arte). Define el conjunto de acciones que debe realizar el agente para conseguir maximizar su recompensa en función su estado, el cuál es percibido a través del entorno. La *policy* constituye el núcleo del agente y nos permite determinar su comportamiento. Estas políticas pueden ser estocásticas.
- **Recompensa (R_t)**. Define el objetivo del agente en un problema de aprendizaje por refuerzo. En cada salto de tiempo (*step*) el agente recibe una recompensa por parte del entorno.
- **Función de valor (V_s^π)**. Representa la máxima recompensa que puede esperar obtener un agente desde un estado concreto, empleando una política concreta, es decir, tiene en cuenta la recompensa a largo plazo, no solo la inmediata.

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s \right] \quad (3.13)$$

Algunos algoritmos de aprendizaje por refuerzo requieren de la definición de un **modelo del entorno**. Este modelo permite predecir el comportamiento que va a tener el entorno a lo largo del tiempo. Los algoritmos que emplean un modelo se les denomina *model-based*, mientras que, a los algoritmos que no requieren de un modelo del entorno se les denomina *model-free*.

Procesos de decisión de Markov

El aprendizaje por refuerzo emplea el marco formal de los procesos de decisión de Markov (*MDP*) en los cuales para definir la interacción entre el agente y el entorno en términos de estados, acciones y recompensas.

Un proceso de decisión de Markov está compuesto por la 4-tupla (S, A, P_a, R_a) donde:

- **S** representa el estado del agente.
- **A** es el conjunto de acciones que puede realizar el agente. A_s denota las acciones que puede realizar el agente desde un estado s .
- $P_a(s, s') = Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$. Partiendo de un estado s , P_a representa la probabilidad de pasar al estado s' tomando la acción a .

- $R_a(s, s')$ denota la recompensa inmediata que recibiría el agente al realizar la transición de s a s' mediante la acción a .

Estos procesos cumplen la propiedad de Markov, es decir, que el pasado no influye en el agente, lo único relevante es el estado actual.

El problema principal de los MDP es encontrar la secuencia de acciones que debe realizar el agente para maximizar la recompensa a largo plazo, es decir, encontrar la política π óptima, que permita maximizar la recompensa. Una vez encontrada la política π óptima el problema se reduce a una cadena de Markov, dado que la acción a a realizar en un estado s viene completamente definida por el estado s y la probabilidad P_a .

La política π óptima es aquella que maximice la recompensa acumulada, es decir, la suma descontada de las recompensas instantáneas percibidas por el entorno:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (3.14)$$

donde γ es el factor de descuento, el cual debe cumplir $0 \leq \gamma \leq 1$, cuanto mayor sea este factor menos importante es la recompensa inmediata.

3.3.1. Algoritmos de *Q-learning*

En este trabajo se han trabajado con 2 tipos de algoritmos de aprendizaje por refuerzo: métodos basados en *Q-learning* y métodos de gradiente de las políticas.

Los algoritmos de Q-learning se basan en la función estado-acción, también conocida como función Q, de ahí el nombre de estos algoritmos. Esta función denota el valor de tomar una acción para un estado concreto, siguiendo una política π . Formalmente, la función Q se define como

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \middle| s_t = s, a_t = a \right] \quad (3.15)$$

La diferencia entre la función Q y la función de valor V^π radica en que la función de valor cuantifica el valor, en términos de recompensa acumulada, que posee un estado concreto, mientras que la función Q evalúa la conveniencia de tomar una acción concreta en un estado determinado. Este matiz se empleará posteriormente para definir la función de ventaja en el apartado 3.3.2.

El objetivo de estos algoritmos se basa en obtener la función óptima Q^* , esta función cumple la ecuación de Bellman

$$Q^*(s, a) = \mathbb{E}_{s'} \left[r + \gamma \max_{a'} Q^*(s', a') \middle| s, a \right] \quad (3.16)$$

siendo s' y a' la secuencia siguiente de estados y acciones, respectivamente.

Basándose en esta ecuación la función Q^* se podría calcular como un proceso iterativo de la forma

$$Q_{i+1}(s, a) = \mathbb{E} \left[r + \gamma \max_{a'} Q_i(s', a') \middle| s, a \right] \quad (3.17)$$

Aunque este proceso converge a la función estado-acción óptima, $Q_i \rightarrow Q^*$ cuando $i \rightarrow \infty$ [9], esto no es práctico, debido a que esta aproximación de la función Q se estima de forma independiente para cada secuencia concreta, sin generalizar. En lugar de esto, se suele emplear una función que se aproxime a esta función estado-acción. En los casos en

los que esta función sea aproximada por una red neuronal, ésta se denota por $Q(s, a; \theta)$ siendo θ los pesos de esta red-Q (*Q-network*).

La función de pérdida que se intenta minimizar para entrenar esta red es

$$L_i(\theta_i) = \mathbb{E}_{s,a \sim \rho(s,a)} [(y_i - Q(s, a; \theta_i))^2] \quad (3.18)$$

donde $y_i = \mathbb{E}_{s'}[r + \gamma \max_{a'} Q(s', a'; \theta_{i-1})]$ es el objetivo(*target*) para la iteración i y $\rho(s, a)$ es la distribución del comportamiento, es decir, distribución de probabilidad de las secuencias a', s' a lo largo del tiempo.

Esta función de pérdida se puede minimizar con algoritmos de optimización como el Descenso de gradientes estocástico (SDG).

DQN

El algoritmo DQN (*Deep Q-network*) es un algoritmo basado en *Q-learning* desarrollado por Mnih et al. [10] en 2015. Con este algoritmo la empresa DeepMind consiguió desarrollar un agente capaz de jugar a los juegos de la clásica consola Atari, superando con creces el rendimiento de los jugadores profesionales en algunos juegos. Este algoritmo está diseñado para actuar sobre un conjunto de acciones discretas, como son las acciones de la consola Atari.

El nombre del algoritmo, hace referencia al empleo de una red neuronal profunda (*Deep Network*) como aproximador de la función Q (*Q-network*). Este algoritmo introduce, principalmente, dos modificaciones sobre los algoritmos Q que favorecen el aprendizaje:

1. **Creación de un “contenedor de repeticiones”** (*buffer replay*) en el que se almacenan las experiencias del agente en cada iteración temporal $e_t = (s_t, a_t, r_t, s_{t+1})$ en un conjunto de datos $D_t = \{e_1, \dots, e_t\}$ recabados durante varios episodios.

Este contenedor se emplea para obtener una muestra de aleatoriedad de experiencias y constituir un pequeño lote con el que optimizar la red-Q.

Emplear este método proporciona varias ventajas notorias:

- Permite el empleo de los datos recabados en cada paso temporal para optimizar los pesos en múltiples interacciones, lo que se traduce en un mejor aprovechamiento de las experiencias del agente.
- Al tomar muestras de forma consecutiva el aprendizaje se vuelve ineficiente debido a la alta correlación entre las muestras. Al tomar muestras aleatorias del contenedor se consigue romper esta correlación.

2. **Empleo de una red objetivo Q'** para la actualización de pesos. Cada C actualizaciones se clona la red Q para obtener la red objetivo \hat{Q} . Esta red objetivo \hat{Q} es la que emplea para realizar el cálculo de la expresión y_i durante la actualización de pesos.

$$y_i = \mathbb{E}_{s'}[r + \gamma \max_{a'} \hat{Q}(s', a'; \theta^-)] \quad (3.19)$$

En este algoritmo se realiza la optimización mediante el algoritmo SDG de la función de pérdida 3.18. Cada C pasos se reinicia la función Q con los valores del objetivo $Q = \hat{Q}$.

Al emplear una red con parámetros desactualizados, a la hora de actualizar la red, se reduce las oscilaciones en el entrenamiento y se favorece la convergencia.

En este algoritmo, la política π que dicta que acción a es la más conveniente en un estado s determinado se obtiene a partir de la función Q de forma inmediata

$$a_t(s) = \operatorname{argmax}_a Q(s, a | \theta) \quad (3.20)$$

es decir, se escoge la acción a que maximice la función Q en ese estado.

DDPG

El algoritmo DDPG (*Deep Deterministic Policy Gradient*), creado por Lillicrap et al. [11] en 2016, adapta las ideas del DQN a un dominio de acciones continuo. Este algoritmo emplea la filosofía *actor-critic*, en la que una parte evalúa el valor de una acción en un estado (crítico) y otra define la política que debe realizar el agente (actor).

La función actor $\mu(s | \theta^\mu)$ devuelve la acción que se debe realizar en el estado s , es decir, la política π que debe seguir el agente. El crítico $Q(s, a)$ se entrena empleando la identidad de Bellman al igual que en los algoritmos de *Q-learning*.

Para actualizar los pesos de la función actor se emplea el gradiente de la función de coste empleados por Silver et al. [12]

$$\begin{aligned} \nabla_{\theta^\mu} J &\approx \mathbb{E}_{s_t \sim \rho} [\nabla_{\theta^\mu} Q(s, a | \theta^Q)|_{s=s_t, a=\mu(s_t | \theta^\mu)}] \\ &= \mathbb{E}_{s_t \sim \rho} [\nabla_a Q(s, a | \theta^Q)|_{s=s_t, a=\mu(s_t)} \nabla_{\theta^\mu} \mu(s | \theta^\mu)|_{s=s_t}] \end{aligned} \quad (3.21)$$

Para favorecer la exploración del algoritmo se genera una política de exploración μ' se le añade ruido \mathcal{N} a la política del actor

$$\mu'(s_t) = \mu(s_t | \theta_t^\mu) + \mathcal{N} \quad (3.22)$$

este ruido se puede generar de varias formas, en el artículo se emplea ruido generado mediante el proceso de Ornstein-Uhlenbeck [13] para generar ruido correlado temporalmente.

Este algoritmo mantiene el contenedor de repeticiones y el empleo de redes objetivo del DQN aunque añade alguna modificación en esta última. En vez de reiniciar el valor de la red con el valor de la red objetivo cada C épocas, en el DDPG, esta actualización de las redes Q y μ se realiza de forma suave mediante una media ponderada móvil.

$$\theta^{\hat{Q}} \leftarrow \tau \theta^Q + (1 - \tau) \theta^{\hat{Q}} \quad (3.23)$$

$$\theta^{\hat{\mu}} \leftarrow \tau \theta^\mu + (1 - \tau) \theta^{\hat{\mu}} \quad (3.24)$$

con $\tau \ll 1$. De esta forma los pesos de la red cambian de forma suave, lo que favorece la estabilidad del entrenamiento.

3.3.2. Algoritmos de gradiente de política

En los algoritmos de *Q-learning* es objetivo es encontrar una función $Q(s, a | \theta)$ que se aproxime lo máximo a la función Q^* , para a partir de esta función extraer la política óptima. En cambio, en los algoritmos de gradiente de política (*policy gradients*) lo que se parametriza, es una política π_θ , por lo que el objetivo de estos algoritmos es encontrar los parámetros que maximizan la recompensa acumulada. La función de coste a maximizar de estos algoritmos es directamente

$$J(\theta) = \mathbb{E}_\pi[r(\tau)] \quad (3.25)$$

siendo $r(\tau)$ la recompensa total obtenida al realizar una trayectoria τ . Si se emplea un algoritmo de optimización basado en el descenso de gradientes los pesos se actualizarian de la siguiente forma

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta_t) \quad (3.26)$$

por lo que el objetivo es encontrar el gradiente $\nabla J(\theta_t)$ e iterar en el tiempo para obtener la política optima. El teorema del gradiente de política [12] se obtiene que

$$\nabla J(\theta_t) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(a_t|s_t) A^{\pi_\theta}(s_t, a_t) \right] \quad (3.27)$$

Siendo A^π la función de ventaja, definida como

$$A^\pi(s_t, a_t) = Q^\pi(s_t, a_t) - V^\pi(s_t) \quad (3.28)$$

intuitivamente, la función de ventaja representa el beneficio o perjuicio de elegir la acción a_t en lugar de seguir la política π en un estado determinado. Al emplear esta función en la expresión 3.27 la política evoluciona hacia las acciones que consiguen una recompensa mayor que la acción media.

TRPO

El algoritmo TRPO (*Trust Region Policy Optimization*) es un algoritmo desarrollado por Schulman et al. [14] en 2017, que emplea el concepto de la región de confianza (*Trust Region*) para favorecer la convergencia del entrenamiento.

La región de confianza limita el cambio que puede sufrir la política en cada iteración, de esta forma, primero se establece el tamaño máximo del paso y posteriormente se localiza el punto óptimo dentro de esta región.

En este algoritmo se intenta

$$\text{Maximizar}_{\theta} \quad \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right] \quad (3.29)$$

$$\text{Sujeto a} \quad \hat{\mathbb{E}}_t [\text{KL}[\pi_{\theta_{old}}(\cdot|s_t), \pi_\theta(\cdot|s_t)]] \leq \delta \quad (3.30)$$

La expresión 3.29 es una función límite inferior a $J(\theta^*)$, es decir, para cualquier valor de $\theta \neq \theta^*$, todos los puntos de la expresión 3.29 se encuentran por debajo de la función $J(\theta^*)$. Esto implica que se pueden emplear métodos de minorización-maximización (algoritmos MM), los cuales van aproximando una función límite inferior, a otra, límite superior mediante iteraciones (fig. 3.10).

Al emplear este método, junto con las limitaciones impuestas con la región de confianza, se consigue mejorar la convergencia del entrenamiento a la solución óptima.

PPO

El algoritmo PPO (*Proximal Policy Optimization*), fue desarrollado por Schulman et al. [14] en 2017, como una mejora del algoritmo TRPO explicado anteriormente. El algoritmo TRPO tiene una complejidad de implementación muy elevada, el PPO simplifica esta complejidad y consigue mejorar el rendimiento de su predecesor.

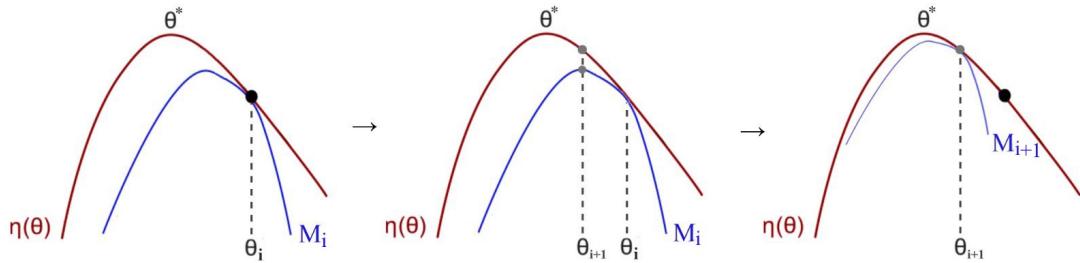


Figura 3.10: Evolución de una función límite inferior con un algoritmo MM.

Este algoritmo tiene dos variantes, referidas en el artículo [14] como *Adaptative KL Penalty Coefficient* y *Clipper Surrogate Objetive*. aunque, en este trabajo, solamente se tratará sobre esta última, la variante que “recorta” el objetivo

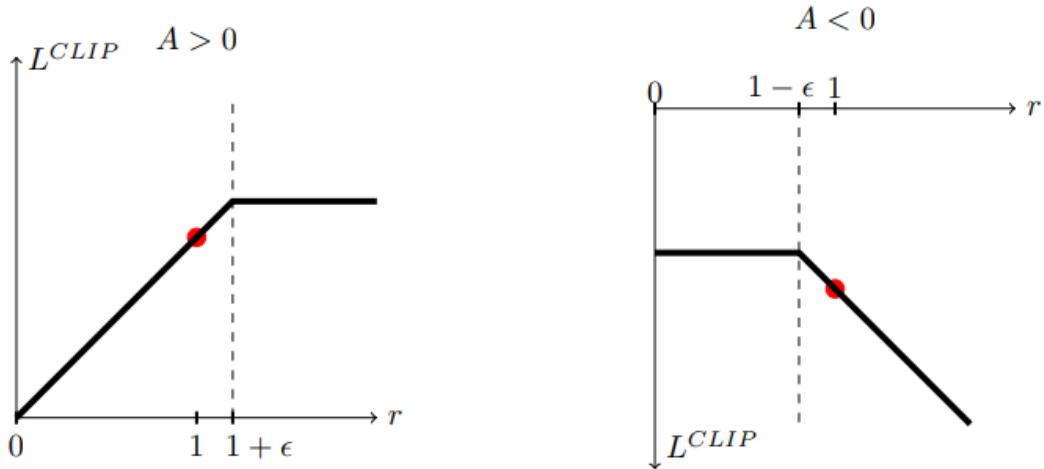
Sea $r_t(\theta)$ el radio de probabilidad $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$, empleando esto en la expresión de la función objetivo del TRPO (eq. 3.29)

$$L^{\text{CPI}}(\theta) = \hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \hat{A}_t \right] = \hat{\mathbb{E}}_t [r_t(\theta) \hat{A}_t] \quad (3.31)$$

Si no hubiera ninguna limitación, al maximizar L^{CPI} , se realizarían cambios muy grandes de la política. Es por esto que Schulman et al. se plantean como modificar esta función para penalizar cambios grandes en las políticas, los cuales distancian $r_t(\theta)$ de 1. La función objetivo que proponen

$$L^{\text{CLIP}}(\theta) = \hat{\mathbb{E}}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right] \quad (3.32)$$

donde ϵ es un hiperparámetro, por ejemplo, $\epsilon = 0,2$. De esta forma, se limita que se potencien mucho las acciones realizadas, cuando la ventaja es positiva y que , cuando la ventaja sea negativa, no se rechacen de forma permanente las acciones realizadas, véase fig. 3.11.

Figura 3.11: Gráficas de la función objetivo $L^{\text{CLIP}}(\theta)$ para distintos valores de r , dependiendo del signo de la ventaja A

De esta manera, la función objetivo se puede optimizar con un optimizador de primer orden, por lo que, aunque a veces se tomen algunas acciones erróneas, se consigue obtener un mejor rendimiento de una forma mucho mas sencilla al TRPO.

Hardware

Con el ánimo de tener una plataforma real sobre la que probar el rendimiento de los algoritmos de control diseñados, se ha diseñado y construido un cuadrcóptero *ad hoc* para este propósito. Un cuadricóptero convencional cuenta con: un chasis o cuadro que lo sustenta, cuatro motores y la electrónica necesaria para controlarlos, una controladora de vuelo que lo comanda y baterías que le proporcionan energía.



Figura 4.1: Cuadricóptero diseñado con el autopiloto incorporado.

A continuación se detallará como son las distintas partes físicas del cuadricóptero que se ha fabricado.

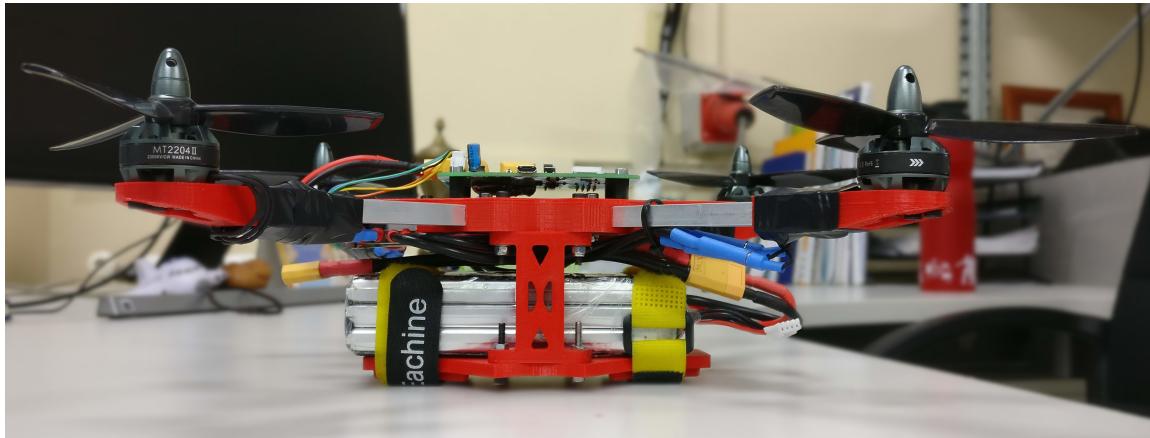


Figura 4.2: Cuadricóptero diseñado con el autopiloto incorporado.

4.1. Cuadro

El *frame* está compuesto por perfiles de aluminio y piezas de PLA fabricadas mediante impresión 3D de diseño propio. La estructura básica está formada por 5 conjuntos de piezas distintas:

- **Portamotores:** son las piezas donde se alojan los motores. Para conseguir una buena fijación los motores se atornillan a los portamotores empleando 4 tornillos dispuestos según los vértices de un rombo.

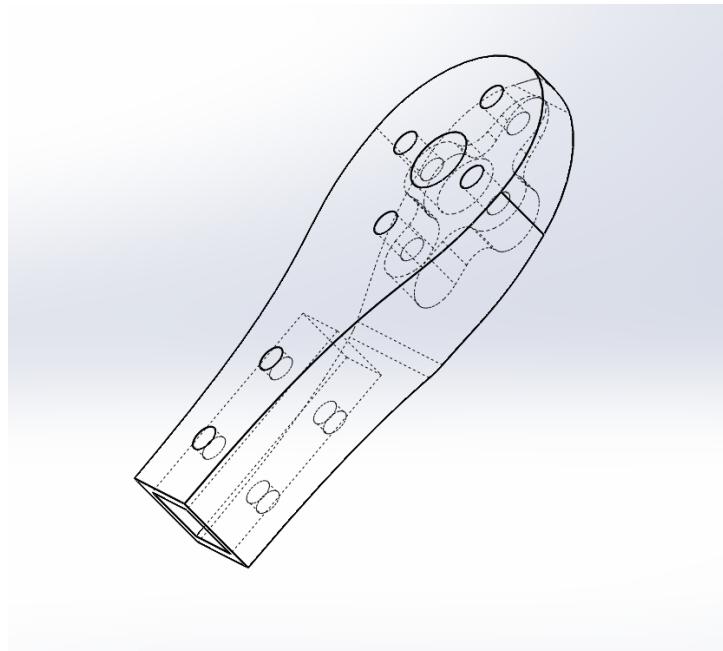


Figura 4.3: Portamotores en CAD

- **Brazos:** se encargan de unir los portamotores con el *núcleo* de la estructura. En este diseño, los brazos consisten en perfiles de aluminio de sección cuadrada de 8mm de lado.
- **Núcleo:** Es la pieza principal del diseño, en la que se anclan el resto de las partes y donde se alojan los componentes electrónicos. Esta pieza sustenta los brazos y el

portabaterías de la aeronave. Cuenta con agujeros a medida para poder situar el autopiloto y los variadores.

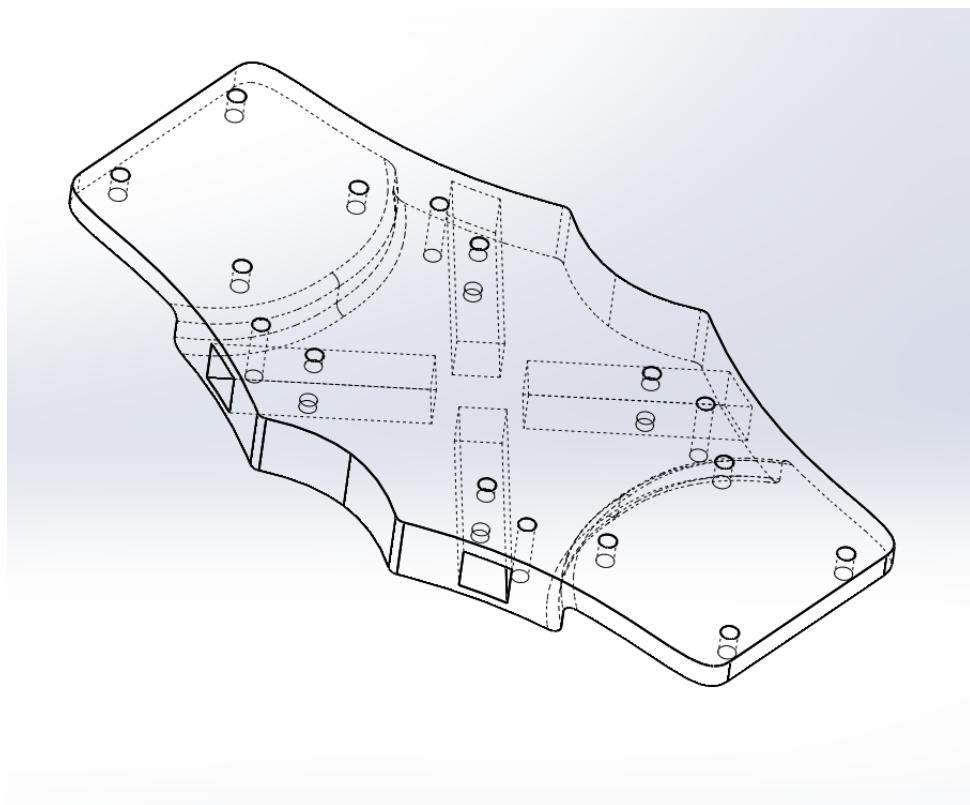


Figura 4.4: Núcleo en CAD

- **Separadores:** su propósito es mantener unidos el *núcleo* y el portabaterías manteniendo una separación fija entre ellas.

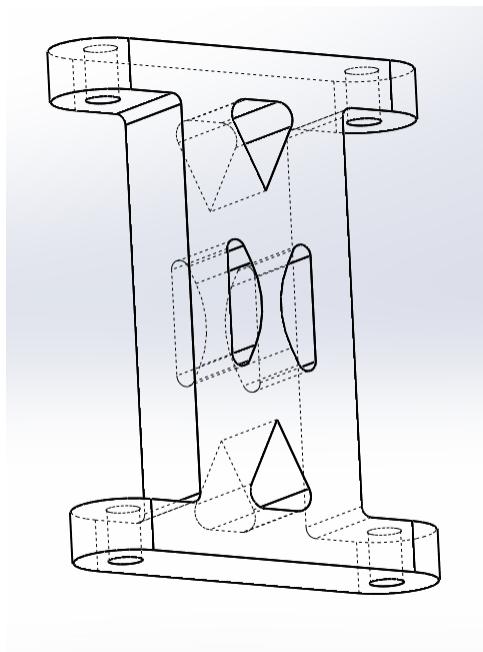


Figura 4.5: Separadores en CAD

- **Portabaterías:** Es la parte inferior del cuadricóptero. En ella se apoya la batería Li-Po que alimenta al cuadricóptero y se mantiene anclada durante el vuelo. Además posee unas protuberancias cuya función se asemejaría a las de un tren de aterrizaje.

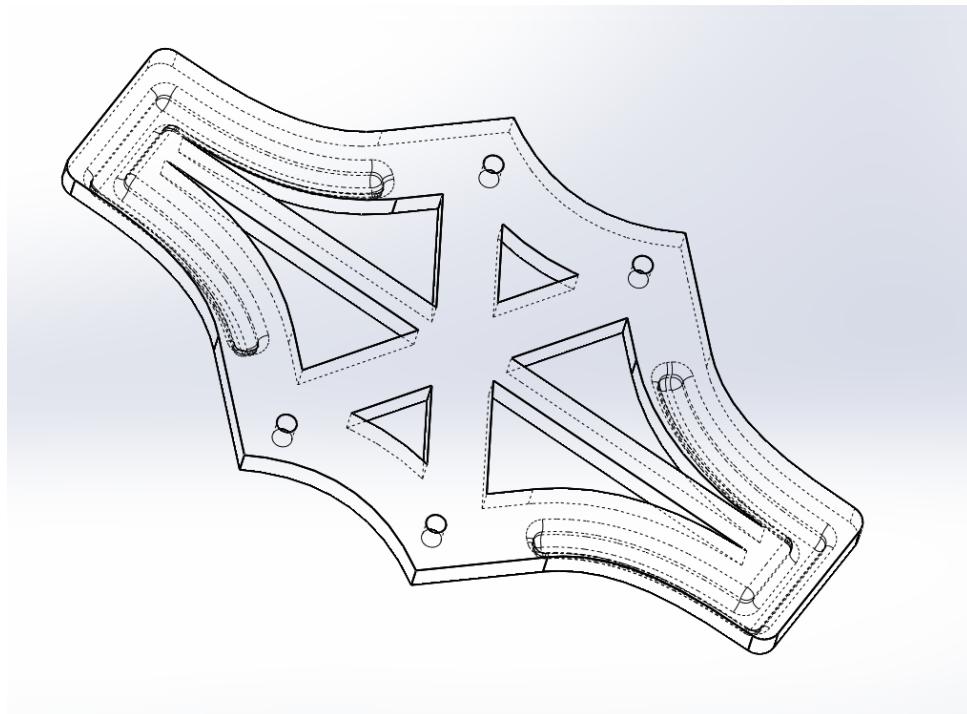


Figura 4.6: Portabaterías en CAD

4.2. Motores y hélices

El cuadricóptero cuenta con 4 motores sin escobillas (*brushless*) LHI MT2204 II de 2300KV con una tensión de alimentación entre 7.2 V y 11.1 V (2s -3s en una batería LiPo) y una corriente continua máxima de 16A.

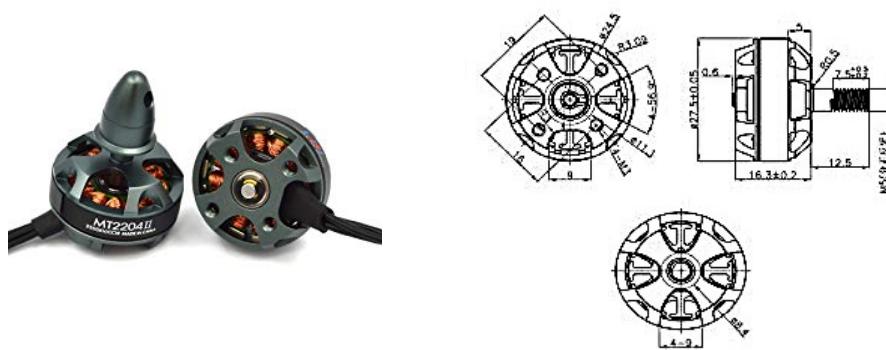


Figura 4.7: Motores LHI MT2204 II empleados

Se ha estimado que el peso aproximado de la aeronave se encuentra entorno a los 800 gramos. La literatura recomienda que los motores que se escojan deben tener empuje suficiente para poder levantar el doble del peso de la aeronave. Observando la tabla de especificaciones de los motores, se observa que, la hélice HQ5040 proporciona empuje

suficiente con un ratio empuje/potencia bastante elevado, como se puede observar en la fig. 4.8.

Motor type	The voltage (V)	Propeller size	current (A)	thrust (G)	power (W)	efficiency (G/W)	speed (RPM)
MT2204 II - 2300KV	8	HQ5040	4.9	210	39.2	5.4	13840
		HQ6045	8.2	320	65.6	4.9	11300
		6030 CF	6.4	240	51.2	4.7	11910
	12	5030 CF	7.5	310	90.0	3.4	20100
		6030 CF	11.5	440	138.0	3.2	16300
		HQ5040	8.4	390	100.8	3.9	19040
		HQ6045	13.2	530	158.4	3.3	14600
	14.8	HQ5040	10.7	510	158.4	3.2	22180
		HQ6045	15.7	620	232.4	2.7	16100

Figura 4.8: Tabla de especificaciones motor MT2204 II.

Es por esto que se han empleado hélices tripala HQ5040 de policarbonato y fibra de vidrio (fig. 4.9).



Figura 4.9: Hélices tripala HQ5040

4.3. Variadores (ESC)

Los motores que se han escogido son trifásicos, es decir, se alimentan con 3 corrientes alternas monofásicas de igual frecuencia y amplitud, desfasadas 120° eléctricos. Para obtener estas formas de ondas a partir de la corriente continua de las baterías, se utilizan los variadores.

Un variador o *ESC* (*Electronic Speed Control*) es un circuito electrónico que se encarga de generar las ondas eléctricas necesarias para controlar y regular la velocidad de un motor eléctrico. Cuenta con un microcontrolador, el cual se encarga de comutar los interruptores de potencia, con el ánimo de alimentar las distintas fases del motor de forma sincronizada, haciendo que gire a la velocidad deseada, véase la fig. 5.2.

Para el cuadricóptero se ha optado por emplear un variador BLHeli Multistar Race (fig. 4.11) que integra 4 variadores en uno, es decir se pueden alimentar 4 motores trifásicos con él. Estos variadores soportan una corriente de hasta 30 A cada uno.

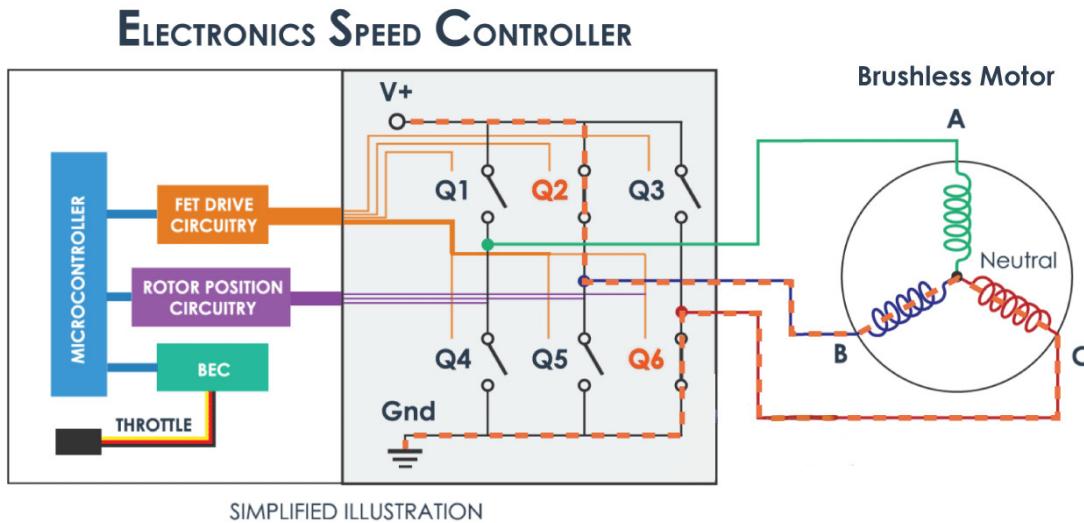


Figura 4.10: Funcionamiento ESC (www.hwtomechatronics.com)

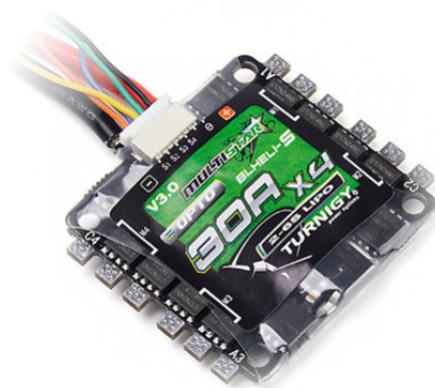


Figura 4.11: ESC Multistar Race 4 in 1 30A BLHeli empleado

4.4. Baterías

Para alimentar al cuadricóptero, se han elegido baterías de litio-polímero (LiPo) de 3 celdas, lo que supone una tensión nominal de 11.1 V (cada celda proporciona 2.7 V). La tensión máxima que puede suministrar la batería es de 12.6 V y la tensión mínima es de 9.6 V. Si la tensión de alguna celda desciende de 3.2 V la batería se puede dañar permanentemente. La batería escogida es una ZNACE de 3 celdas, con una capacidad de 5200 mAh y una tasa de descarga de 35 C, por lo que es capaz de proporcionar una corriente de salida máxima de $5.2 \text{ Ah} \cdot 35 \text{ C} = 182 \text{ A}$. La batería empleada cuenta con un conector XT60, por lo que el máximo de corriente que soporta en régimen continuo es de 60 A.

4.5. Autopiloto

En los cuadricópteros, el sistema que se encarga de estabilizarlo y hacerlo pilotable se denomina la controladora de vuelo o el Autopiloto. Existe una gran variedad de controladoras de vuelo, pero las más populares son las basadas en software abierto como ArduPilot y Betaflight.



Figura 4.12: Batería LiPo 3s 35C 5200 mAh de la marca NZACE empleada

doras en el mercado, pero para este trabajo se ha diseñado una controladora propia con el fin de poder tener acceso a todos los sensores y a implementar el algoritmo de control de forma óptima. El autopiloto consta de 3 partes diferenciadas: la electrónica de potencia, el microcontrolador y los sensores. A continuación se tratará sobre estas partes con más detalle.

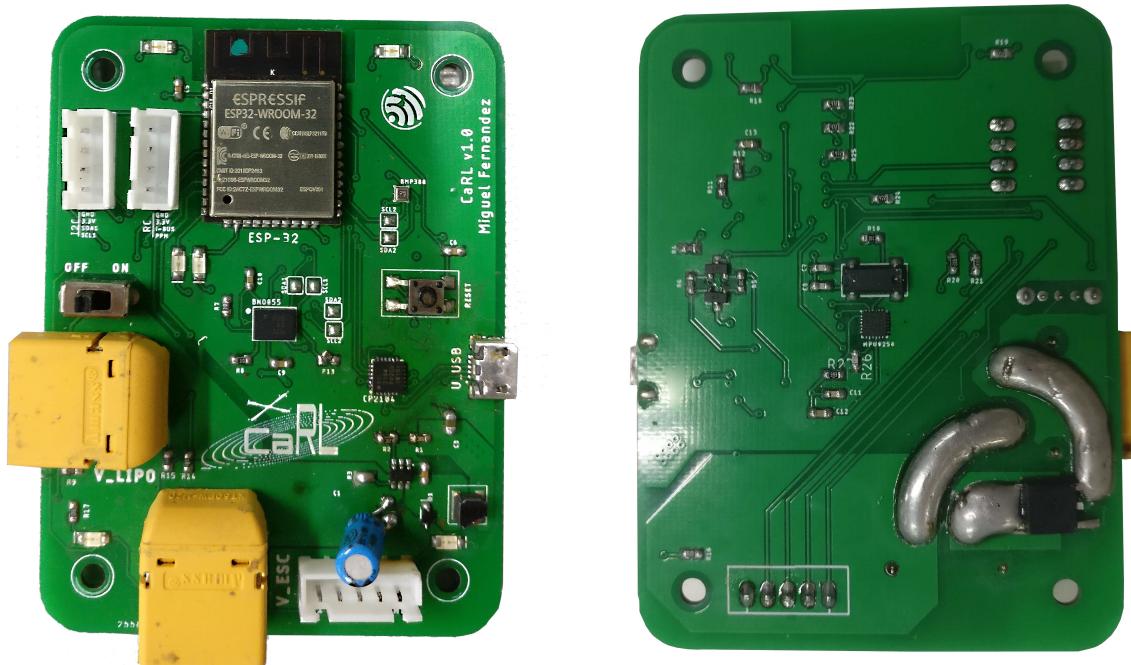


Figura 4.13: PCB autopiloto CaRL, anverso y reverso.

4.5.1. Fase de Potencia

Con el fin de poder gestionar la potencia entregada por las baterías a la placa y a los motores se ha diseñado una etapa de potencia en la que se debe mencionar dos partes: el interruptor de potencia y el regulador a 3.3 V.

Interruptor de potencia

Los motores del cuadricóptero pueden llegar a consumir 12 A cada uno, lo que los cuatro motores pueden llegar a consumir 48 A. Un interruptor con tamaño reducido no puede manejar tanta corriente, por ello se ha empleado un transistor MOSFET de canal P por el que pueden circular hasta 100 A, con el fin de abrir o cerrar el paso de corriente desde las baterías al resto de la placa. El MOSFET se controla con un interruptor de poca potencia entre drenador y puerta. Cuando se cierra el interruptor se alimenta directamente al ESC y al regulador de tensión.

Adicionalmente, a continuación del MOSFET se encuentra un divisor de tensión para poder medir la tensión de la batería a través del conversor analógico digital (ADC) del microcontrolador.

Regulador a 3.3V

La electrónica digital de la PCB se alimenta y emplea lógica a 3.3 V, por lo que no la podemos conectar a las baterías de 11.1 V. Para adecuar la tensión se ha escogido un regulador Step-down de tipo Buck (Figura 4.14). El circuito integrado que se encarga de conmutar la fuente es el chip AP3211.

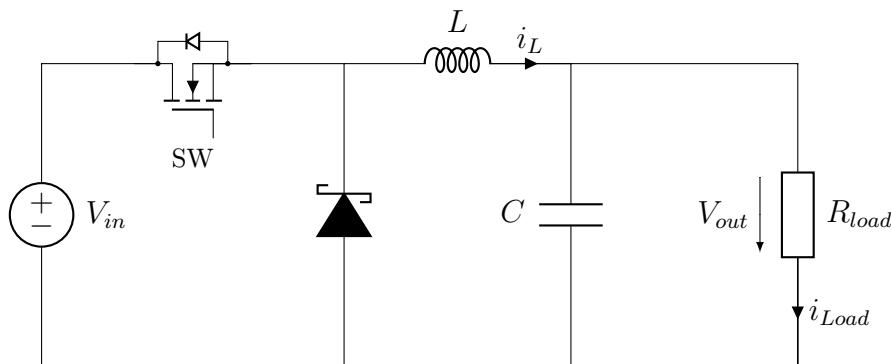


Figura 4.14: Esquema de un convertidor Buck

4.5.2. El microcontrolador (ESP32)

El microcontrolador por el que se ha optado para este Autopiloto es el ESP32, un microcontrolador de doble núcleo con dos CPUs XTensaL6 con arquitectura Harvard [15]. El ESP32 tiene una frecuencia de reloj de hasta 240MHz ,y cuenta con una antena WiFi a 2,4 GHz y conexión Bluetooth 4.2 BLE [16]. Los motivos por los que se ha decidido emplear este microcontrolador son:

- Elevada frecuencia de procesamiento y dos núcleos de procesamiento.
- Antena WiFi incorporada.
- Bajo consumo de potencia.

Para poder programar el microcontrolador se utiliza un convertidor USB (Bus Serie Universal) a UART (Transmisor-Receptor Asíncrono Universal) que permite conectar por USB el microcontrolador para poder programarlo y hacer depuración utilizando comunicaciones Serial. El chip que realiza esta función es el CP2104.

4.5.3. Sensores

La principal fuente de información procedente del exterior que recibe una controladora de vuelo se la proporcionan las unidades de medición inercial (IMU). Las IMUs son dispositivos electrónicos que son capaces de medir aceleraciones lineales, velocidades angulares y detectar la orientación de un sistema. El principal problema de estos sensores es que sufren error acumulativo a la hora de estimar posición y velocidad. Para corregir este error acumulativo en los cuadricópteros, se suelen fusionar estas medidas con otras provenientes de mediciones absolutas tales como GPS o Láser, aunque en este autopiloto únicamente emplearemos las medidas de las IMUs.

Otros sensores utilizados frecuentemente en los autopilotos son brújulas (se encuentran integrados en la IMU para corregir errores de orientación) y barómetros (para estimar la altitud a la que se encuentra el cuadricóptero).

El autopiloto cuenta con dos IMUs de 9 Grados de Libertad y un barómetro para conseguir una mejor estimación del estado del cuadricóptero:

1. **BNO 055 (BOSCH)**: El circuito integrado de Bosch es un sensor “inteligente” que incluye los sensores y la fusión de las lecturas de los distintos sensores en un único componente. Este encapsulado cuenta con: un acelerómetro, un giróscopo y un magnetómetro triaxial. Además integra un microcontrolador de 32 bits en el que se ejecuta el algoritmo de fusión integrado. El sensor se encuentra en un encapsulado LGA de 28 pines con una huella (*footprint*) de $3,8 \times 5,2 \text{ mm}^2$.

Este sensor nos proporciona estimaciones del estado completo de la aeronave con una frecuencia de refresco de 100Hz.

2. **MPU 9250 (TDK InvenSense)**: El sensor inercial de TDK es un módulo multi-chip compuesto por un MPU6050 (Contiene un acelerómetro y un giróscopo triaxial con un *procesador digital de movimiento* (DMP)) y un AK8963 (un magnetómetro digital triaxial). El sensor posee un encapsulado QFN de $3 \times 3 \times 1 \text{ mm}$ con 24 pines.

Este dispositivo nos proporciona medidas del acelerómetro y el giróscopo a una frecuencia superior al BNO055 pero con una estimación peor de los ángulos que el dispositivo de BOSCH.

3. **BMP388 (BOSCH)**: El BMP388 es un barómetro digital de 24 bits con bajo consumo y bajo ruido. Se encuentra en un encapsulado LGA de 10 pines de dimensiones $2 \times 2 \times 0,75 \text{ mm}^3$.

Aunque el autopiloto cuenta con este sensor, éste no se ha utilizado en este trabajo debido a que no se tiene en cuenta la altitud en los controladores.

4.6. Banco de pruebas

Para poder realizar la experimentación real de forma segura, se han diseñado distintas estructuras para poder sujetar al cuadricóptero, permitiéndole rotar con distintos grados de libertad (GdL) en función de la estructura. Estas uniones han permitido poder probar distintos controladores de forma segura y controlada.

Se pueden distinguir 2 tipos de estructuras en función de sus grados de libertad:

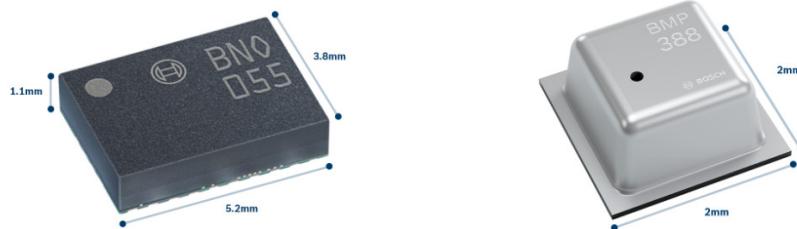


Figura 4.15: Sensores BNO055 y BMP388 respectivamente

- **Rótulas con 1 único grado de libertad:** Para las primeras pruebas de los reguladores es fundamental poder descomponer el control en problemas más sencillos, en este caso permitiendo al sistema rotar en 1 grado de libertad. Se han construido 2 versiones de la misma rótula, una permite el movimiento en Pitch y la otra lo permite en Roll. Esto permite desacoplar los bucles de control y poder probar distintos algoritmos de control.

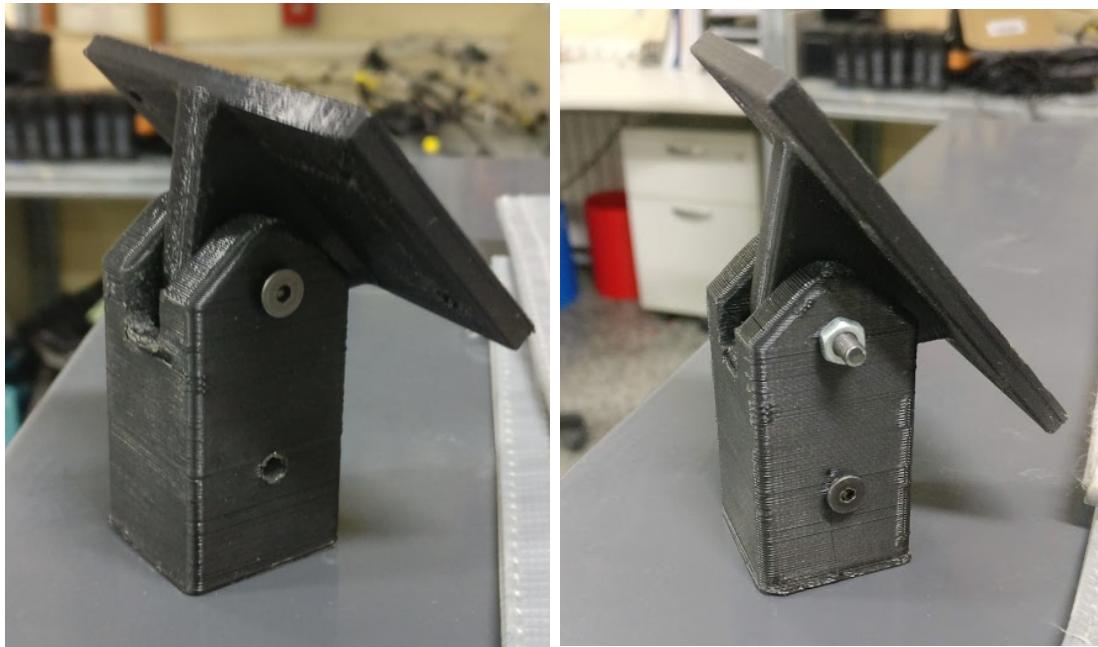


Figura 4.16: Rotulas de 1GdL (*pitch* y *roll* respectivamente)

Estas uniones son sencillas y robustas lo que nos da seguridad a la hora de poder probar y ajustar los reguladores. Las restricciones de movimiento de estas uniones permiten rotaciones de $\pm 60^\circ$ en el angulo permitido.

- **Rótula con múltiples grados de libertad:** Después de conseguir estabilizar al cuadricóptero en *pitch* y en *roll* de forma individual la siguiente aproximación consiste en emplear rotulas con 3 GdL. Sobre estas uniones también se han realizado 2 versiones. La primera consta de una única rotula con sus 3 grados de libertad con restricciones de movimiento de:

$$\begin{aligned} -60^\circ \leq \varphi \leq 60^\circ & \quad \text{Roll} \\ -60^\circ \leq \theta \leq 60^\circ & \quad \text{Pitch} \\ -180^\circ \leq \psi \leq 180^\circ & \quad \text{Yaw} \end{aligned}$$



Figura 4.17: Junta esférica antes de ensamblar

La segunda consta de acoplar 2 juntas esféricas como las anteriores una a continuación de la otra, lo que permite, además de disminuir las restricciones de movimiento en las rotaciones, pequeños desplazamientos en el espacio tridimensional.



Figura 4.18: Junta esférica doble

Software

Durante el transcurso de este trabajo se ha dividido el desarrollo del software en varias partes, en las cuales se profundizará a continuación.

5.1. Software del Autopiloto

El autopiloto es la parte del multirrotor que se encarga de generar las acciones de control o comandos, que permitan que el cuadricóptero llegue a un determinado estado. Para generar esas acciones de control, el autopiloto estima el estado de la aeronave en cada instante mediante las lecturas de las IMUs y en función del estado genera las acciones de control pertinentes para mantener la estabilidad de la aeronave. Además el autopiloto que se ha desarrollado cuenta con un interfaz WiFi, por lo que permite enviar y recibir datos del autopiloto a una estación de tierra. A continuación se profundizará en como se han llevado a cabo estas tareas.

5.1.1. Estimación del Estado

Para estimar el estado, el autopiloto toma las medidas procedentes de las 2 IMUs a través del protocolo I2C con una frecuencia de comunicación de 400KHz. Las lecturas del BNO055 proporcionan un estado completo de la aeronave con una frecuencia de refresco de 100Hz. Sin embargo, el MPU9250 nos proporciona medidas relativas a las velocidades angulares y aceleraciones lineales con una tasa de refresco más elevada. Es por esto que se han empleado ambos sensores para conseguir el estado deseado con la mayor precisión y menor tasa de refresco posible.

Con este método se consigue estimar el estado deseado:

$$S_t = (\underbrace{\varphi, \theta, \psi}_{\text{BNO}}, \underbrace{\dot{\varphi}, \dot{\theta}, \dot{\psi}}_{\text{MPU}}) \quad (5.1)$$

consiguiendo minimizar la deriva de la estimación, empleando las lecturas muy estables del BNO055, sin perder la reactividad de las medidas que proporciona el MPU9250.

5.1.2. Interfaz WiFi

Con la intención de poder transmitir datos entre el autopiloto y la estación de tierra se ha diseñado un sencillo protocolo de comunicación WiFi basado en el envío de paquetes de tamaño fijo. El ESP32 cuenta con un sistema operativo en tiempo real (RTOS) el cual se encarga de mantener activa la comunicación WiFi de forma transparente al usuario. La estructura de los paquetes es la siguiente:

- **Autopiloto → Estación:** Este paquete contiene 6 datos de tipo *float* (32-bits). Se emplea para enviar la estimación del estado a la aeronave en tiempo real. Esto se emplea para tanto monitorización del desempeño de los algoritmos, como para poder cerrar el bucle de control en la estación base y enviar las acciones de control a la aeronave posteriormente.
- **Autopiloto ← Estación:** Este paquete contiene 4 datos de tipo *float* (32-bits) y un dato de tipo *byte*. El dato de tipo *byte* se emplea para controlar el modo de funcionamiento de la aeronave, mientras que los otros 4 datos tienen distintos usos en función del modo:
 - **MODO 0 (Desarmado):** La aeronave se encuentra “desarmada” por lo que los motores no reciben acciones de control. En este modo los demás datos del paquete son irrelevantes.
 - **MODO 1 (*Off-board*):** Las acciones de control son emitidas por la estación de tierra, es decir, la aeronave envía la estimación del estado al ordenador y envía el comando recibido a los variadores. En este modo los datos de tipo *float* del paquete representan los comandos de los motores.
 - **MODO 2 (*On-board*):** El algoritmo de control se ejecuta en la aeronave y se generan las señales de control necesarias a una frecuencia fijada previamente. En el caso de los controladores PID a bordo, los valores de los 4 datos del paquete permiten variar los valores de las constantes del regulador en tiempo real, lo que facilita el ajuste de las mismas.

5.1.3. Generacion de comandos *on-board*

Para poder maximizar la frecuencia del algoritmo de control y permitir el control de la aeronave sin necesidad de una estación de tierra, los algoritmos de control validados en la simulación se pueden implementar directamente en el autopiloto, fijando previamente la frecuencia a la que se desea que se ejecute el algoritmo. En este trabajo se han implementado a bordo dos algoritmos de control distintos: un regulador PID clásico y un regulador PID en cascada.

5.2. Software de la estación

La estación base se ha empleado tanto como para la simulación del entorno, permitiendo diseñar y probar algoritmos de control de forma segura, como para monitorizar los rendimientos de estos algoritmos o incluso, aprovechando la mayor capacidad computacional de estas estaciones, ejecutar los algoritmos de control y aprendizaje por refuerzo en estas. A continuación se profundizará con más detalle en la implementación de estas herramientas.

5.2.1. Entorno de simulación

En cualquier proyecto de robótica es conveniente contar con un entorno de simulación que permita realizar pruebas en distintas situación de forma sencilla y rápida, sin necesidad de tener ni modelo real, ni el entorno concreto en el que quieras probar tu robot. Cuando se trabaja con cuadricópteros, la simulación pasa de ser conveniente a ser muy necesaria.

Ésto es debido a la peligrosidad intrínseca de estas aeronaves, ya que, cuentan con hélices que giran a gran velocidad.

Debido a la naturaleza del aprendizaje por refuerzo, este requiere de la interacción de un agente con un entorno para que el agente aprenda que acciones son las que debe tomar en cada estado. Esto significa que al comienzo del entrenamiento el agente realiza acciones aleatorias para poder explorar cuales son las que le proporcionan una recompensa mayor.

Debido a esta forma de explorar, el agente requiere de una gran cantidad de pruebas, de ensayo y error, hasta que consigue aprender, por lo que no es conveniente realizar todas estas iteraciones en un modelo real.

Por estas razones se necesita de un entorno de simulación para poder validar los algoritmos y poder generar modelos entrenados en simulación sin deteriorar el equipo real. Además el entorno en simulación te permite entrenar distintos agentes simultáneamente y reducir los tiempos de entrenamiento.

Para el entorno de simulación nos hemos basado en Gym [17], una librería escrita en Python y desarrollada por la compañía OpenAI, que permite desarrollar y comparar algoritmos de aprendizaje por refuerzo. Esta librería te permite generar un entorno con el cual interactúe el agente, sin importar la implementación de éste, permitiendo comparar el rendimiento de distintos agentes sobre el mismo entorno.

Como entorno de simulación se ha partido de GymFC, desarrollado por William Koch et al. [7]. Este entorno de simulación utiliza Gazebo 9, un entorno de simulación 3D de código abierto ampliamente utilizado en el campo de la robótica. En este entorno se simula el comportamiento de un multirrotor, concretamente el de un modelo del cuadricóptero IRIS.

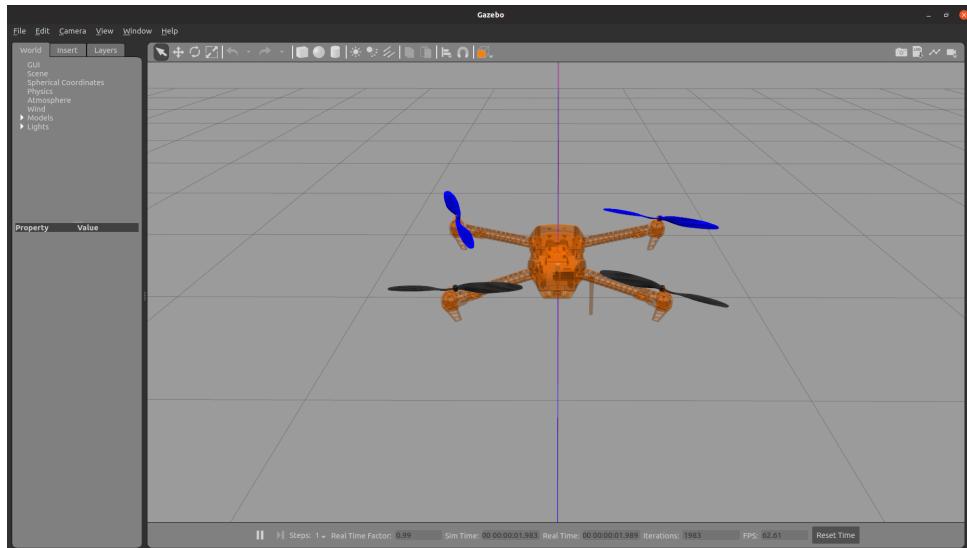


Figura 5.1: Entorno de simulación GymFC en Gazebo 9

Para acercar la simulación a la configuración de los experimentos que se realizarían posteriormente en la plataforma real, se han realizado algunas modificaciones sobre el modelo predeterminado de la aeronave:

1. Se ha modificado la forma de anclaje del cuadricóptero. Originalmente el cuadricóptero se encontraba anclado entorno a una articulación situada en su centro de gravedad. Esto permitía que el peso de la aeronave apenas tuviera influencia a la hora de rotar la aeronave, todo el peso estaba sustentado por la articulación, lo cual

no se asemeja con los bancos de pruebas que se han diseñado. En estos bancos de prueba el anclaje se encuentra desplazado con respecto al centro de gravedad, por lo que cuanto mayor sea el valor de los ángulos φ y θ mayor es la influencia del peso en el par necesario para estabilizar la aeronave. Es por esto que se ha desplazado la articulación del centro de gravedad y se ha conectado a la aeronave mediante una unión rígida, asemejando así el entorno de simulación al entorno de pruebas real.

2. Se ha modificado parámetros dinámicos de la aeronave. Los parámetros que mayor discrepancia tenían entre la simulación y el mundo real eran principalmente los parámetros inerciales. Tanto la masa como los momentos de inercia del cuadricóptero eran mucho inferiores a los de la aeronave real. Esto se manifestaba principalmente cuando al hacer pruebas con los algoritmos PID el comportamiento simulado no se parecía con el real, por ejemplo en simulación un regulador P era capaz de estabilizar al sistema, mientras que en la realidad un regulador P era inestable de por sí y requería de las acciones derivativa e integral para conseguir estabilizarlo.

5.2.2. Agente (generación de comandos)

Siguiendo con la terminología del aprendizaje por refuerzo se va a denominar agente al encargado de generar las acciones de control. Dependiendo del algoritmo de control que se use, se dispone de distintos tipos de agente.

Agente con algoritmos PID

En este caso, el agente recibe las observaciones del entorno a través del estado y en función de éste, produce la salida a los motores. Este agente es el más sencillo, ya que, una vez ajustadas las ganancias, éste solamente debe enviar los comandos de control a la aeronave.

Agente con algoritmos de RL

Este agente, se encarga, tanto de generar las acciones como de optimizar la política que las genera. El agente, basado en el estado observado s_t y la política π actual, genera una acción a_t , tras realizarla, el entorno le proporciona un estado s_{t+1} y una recompensa r_t . Con esta tupla $[s_t, a_t, r_t, s_{t+1}]$ el agente emplea el algoritmo de aprendizaje por refuerzo definido para actualizar la política π y así intentar mejorarlala con cada iteración.

Para probar distintos algoritmos se han empleado la implementación de los algoritmos del estado del arte, del repositorio de GitHub *Stable-Baselines* [18]. Inicialmente se comenzó empleando las *Baselines* de OpenAI [19] pero se decidió cambiar a las *Stable Baselines* por la limpieza del código y la facilidad para realizar experimentos con diversos algoritmos e hiperparámetros. Con esta herramienta se ha podido experimentar con distintos algoritmos para comparar el rendimiento conseguido de cada uno.

5.2.3. Interfaz Estación-Autopiloto

Para las pruebas reales se ha empleado ROS (Robotic Operative System) Melodic para gestionar la comunicación WiFi con la aeronave y la comunicación con el agente. Para esto se emplea la estructura de *topics*: un *topic* se emplea para enviar datos a la aeronave y otro *topic* publica los datos recibidos por ésta. De esta manera se pueden

utilizar las herramientas de ROS para poder monitorizar el proceso además de aprovechar la estructura de *topics* para poder integrar código escrito en distintos lenguajes de forma sencilla.

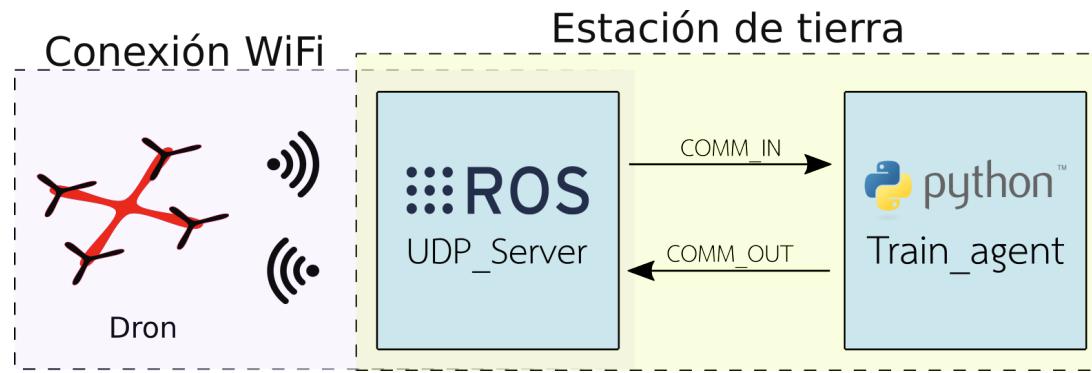


Figura 5.2: Esquema interfaz Estacion-Autopiloto

5.3. Descripción del equipo

Para el desarrollo del trabajo se ha empleado un ordenador portátil MSI-GE62 empleando Windows 10 para el desarrollo CAD y Ubuntu 18.04 LTS para el resto de las tareas. El equipo cuenta con 16GB de RAM DDR4, un procesador Intel i7-6700HQ de 8 núcleos a 2.60GHz y una GPU GeForce GTX 970M con 3GB de memoria dedicada.

Además se ha empleado un ordenador de sobremesa montado por piezas con Ubuntu 18.04. Este equipo cuenta con un procesador Intel i7-9900HQ ,32 GB de RAM DDR4 y una GPU Nvidia EVGA RTX 2080 de 8GB de memoria dedicada.

Metodología

Además del desarrollo de la plataforma de la aeronave, otro objetivo del trabajo es intentar estabilizar un UAV usando algoritmos de control basados en algoritmos de aprendizaje automático. En este apartado, se tratará sobre la metodología que se ha llevado a cabo durante el proceso de desarrollo de los algoritmos. Los principales componentes que intervienen en el agente son el estado, las acciones y la recompensa, para cada problema hay un conjunto que estados, acciones y funciones de recompensa que pueden llevar a que el agente aprenda.

6.1. Diseño del estado

El autopiloto cuenta con 2 IMUs para poder obtener datos sobre su estado. Se quiere estabilizar el cuadricóptero en una orientación concreta, por lo tanto el estado que se ha diseñado consta de 6 parámetros:

$$S = (\varphi, \theta, \psi, \dot{\varphi}, \dot{\theta}, \dot{\psi}) \quad \varphi, \theta, \psi, \dot{\varphi}, \dot{\theta}, \dot{\psi} \in [-1, 1] \quad (6.1)$$

Siendo φ , θ y ψ los ángulos de alabeo (*roll*), cabeceo (*pitch*) y guiñada (*yaw*) del cuadricóptero y $\dot{\varphi}$, $\dot{\theta}$ y $\dot{\psi}$ sus respectivas velocidades. Para favorecer la convergencia del aprendizaje, se ha normalizado el estado para que todas sus componentes estén comprendidas dentro del intervalo $[-1, 1]$.

Los ángulos proporcionan información sobre el estado actual y la velocidad angular sobre los estados pasados y los posibles estados futuros, es decir, proporciona cierta información temporal.

6.2. Diseño de las acciones

Al trabajar con un cuadricóptero podemos actuar sobre la potencia que se le entrega a los motores, por lo que cada acción que realice el agente constará de 4 campos:

$$A = (T_1, T_2, T_3, T_4) \quad T_i \in [-1, 1] \quad (6.2)$$

Siendo T_i la potencia (*Thrust*) normalizada entregada a cada motor. Un valor de $T_1 = -1$ significa que el motor 1 estaría girando a la mínima potencia permitida y un valor de $T_1 = 1$ corresponde a que el motor 1 estaría girando a la máxima potencia.

Las señales de control que admiten los variadores es una señal PWM de 50Hz cuyo ancho de pulso debe estar entre 1ms y 2ms. Un ancho de pulso de 1ms se corresponde con la mínima velocidad el motor y un ancho de 2ms con la máxima, véase fig. 6.1.

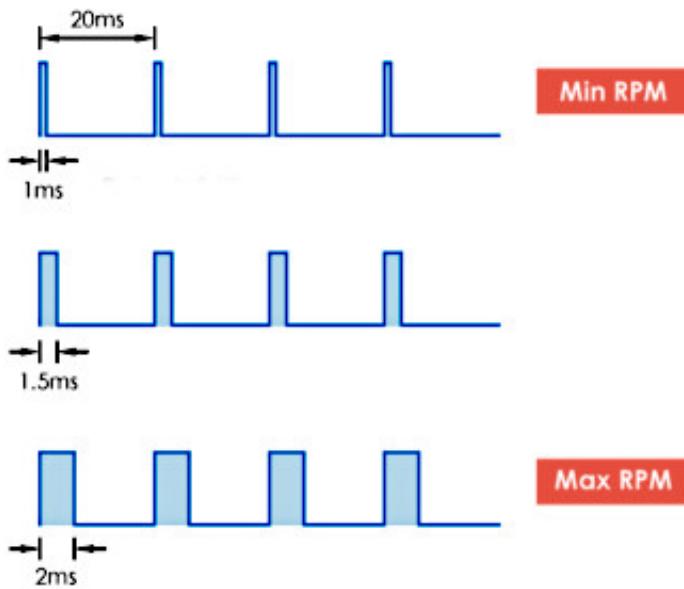


Figura 6.1: Formas de onda que recibe un variador

Para realizar esta transformación del dominio $[-1, 1]$ al dominio $[1000, 2000] \mu\text{s}$ se han empleado la siguiente expresión

$$P_i = T_i \cdot 1000 \cdot \alpha + \beta \quad i = 0, \dots, 3 \quad (6.3)$$

siendo P_i el ancho del pulso enviado al ESC, $\alpha \in [0, 1]$ es el porcentaje de la potencia total que puede manejar el controlador y $\beta \in [0, 2000]$ es la velocidad base que tienen los motores. Esta velocidad base permite que el cuadricóptero pueda variar su altura y que se pueda autosustentar.

6.3. Diseño de la función de recompensa y ajuste de hiperparámetros

La búsqueda de la función de recompensa y la elección de los hiperparámetros que aseguraban la convergencia de los distintos algoritmos, se ha realizado mediante un proceso cíclico, en el que, se realizaba una hipótesis sobre la posible forma de la función de recompensa o el posible valor de un hiperparámetro que podría mejorar el algoritmo. Posteriormente se ejecutaba el algoritmo y se comparaban los resultados de este entrenamiento con los resultados previos. Para disminuir el tiempo de los entrenamientos, se simplificaba el problema del control del cuadricóptero a un único eje de giro. Con esto se redujo el tiempo que se tardaba en completar el ciclo.

Función de recompensa

La función de recompensa rige la forma en la que la red va a configurar sus pesos, por lo tanto, cómo se va a comportar el agente en un estado determinado. Para conseguir que el agente responda de la forma deseada se han probado una gran variedad de funciones

de *reward*, optando finalmente por:

$$R_t = \left(1 - \frac{|\varphi| + |\theta| + |\psi|}{3}\right)^n \quad (6.4)$$

Sea $\gamma = \frac{|\varphi| + |\theta| + |\psi|}{3}$ entonces $R_t = (1 - \gamma)^n$, con $n \in \mathbb{N}$. Aumentando el valor de n podemos conseguir funciones de recompensa con pendientes más pronunciadas.

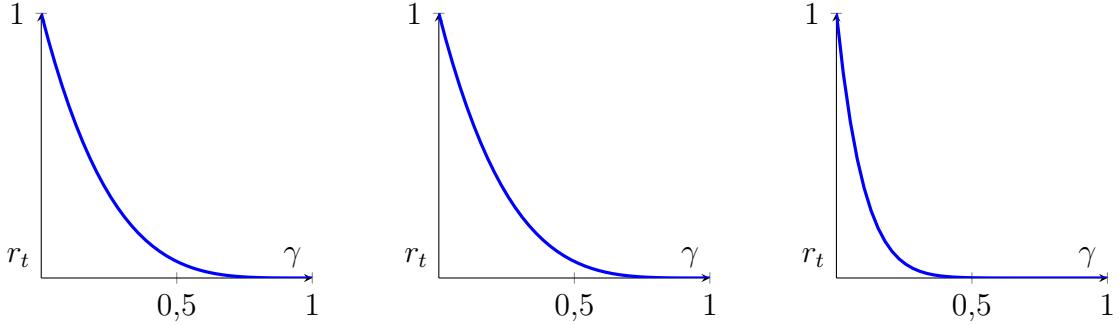


Figura 6.2: Funciones R_t para distintos valores de $n = 2, 4, 10$ respectivamente

Esta función $R_t : [0, 1] \rightarrow [0, 1]$ $\forall n \in \mathbb{N} > 0$ por lo que al proporcionar valores entre 0 y 1, favorece la velocidad del entrenamiento. Además toda la recompensa es positiva, por lo que el agente intentará mantenerse la mayor cantidad de tiempo posible sin reiniciar el episodio, esto es muy importante para conseguir la convergencia del entrenamiento en los escenarios en los que se desea que el cuadricóptero se encuentre siempre dentro de una región concreta del espacio.

Ajuste de hiperparámetros

Debido a las diferentes naturalezas de cada uno de los algoritmos que se han empleado: DDPG, TRPO y PPO , cada uno cuenta con un conjunto distinto de hiperparámetros que ajustar. La implementación de estos algoritmos que se ha empleado (*stable-baselines*) contiene unos hiperparámetros por defecto, los cuales suelen hacer que los algoritmos converjan.

Por lo general, no ha sido necesario modificar mucho estos parámetros por defecto, a excepción del valor de la constante ϵ en el algoritmo PPO. Este parámetro modifica el tamaño de la región de confianza del algoritmo, acotando el tamaño de los pasos que se toman en el curso de cada actualización de la política. Se observó que el valor por defecto de este parámetro $\epsilon_{defecto} = 0,2$ era demasiado grande, por lo que la política divergía. A medida que este parámetro se disminuye, se aumenta la estabilidad del aprendizaje, sin embargo, también se ralentiza mucho. Es por esto que el valor de éste hiperparámetro se ha ido modificando a lo largo de los experimentos, para poder conseguir el entrenamiento más rápido, capaz de converger de forma estable.

Experimentos

Debido a que el trabajo tiene dos partes principales, claramente diferenciadas, se han diseñado distintos tipos de experimentos para la evaluación de cada parte: un conjunto de experimentos en simulación para probar el rendimiento de los algoritmos de aprendizaje por refuerzo y un conjunto experimentos en real para probar la plataforma hardware y el diseño del autopiloto.

7.1. Experimentos en simulación

Con el ánimo de probar los distintos algoritmos de control basados en aprendizaje por refuerzo, se han diseñado diversos experimentos para comparar el rendimiento de estos algoritmos con el de un PID clásico.

Control en 1 GdL (*Roll*)

En este experimento se comparan en rendimiento de los algoritmos PPO, DDPG , TRPO y PID para estabilizar la aeronave en *roll*, restringiendo el movimiento de los otros ejes.

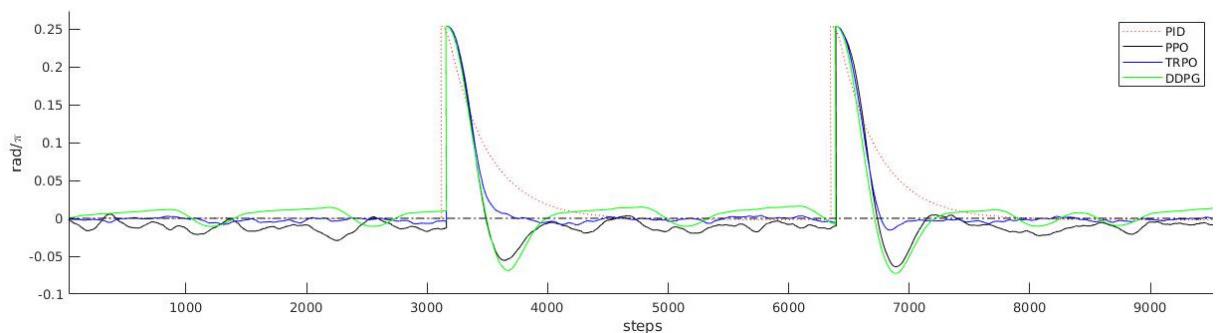


Figura 7.1: Estabilización en *roll* en un entorno simulado.

En este experimento, se observa como el TRPO consigue una mejor respuesta que el resto de los algoritmos. El PPO y el DDPG generan comportamientos similares.

Control en 1 GdL (*Pitch*)

En este experimento se comparan en rendimiento de los algoritmos PPO, DDPG , TRPO y PID para estabilizar la aeronave en *pitch*, restringiendo el movimiento de los otros ejes.

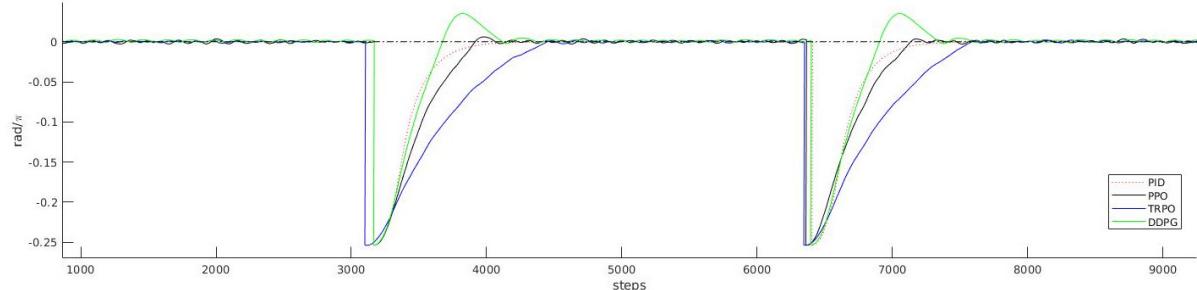


Figura 7.2: Estabilización en *pitch* en un entorno simulado.

En este experimento, se observa como el PPO consigue una mejor respuesta que el resto de los algoritmos. En este entrenamiento el TRPO consigue un peor rendimiento frente a los demás.

Control en los 3 GdL

En este experimento se comparan en rendimiento de los algoritmos PPO, DDPG , TRPO y PID para controlar la estabilización completa en los 3 ejes.

En este experimento, se observa como el TRPO consigue una mejor respuesta que el resto de los algoritmos, mostrando una salida mucho más rápida . El PPO sobreoscila demasiado en *roll* y presenta un error de posición, véase en la Fig. 7.3

7.2. Experimentos en real

Para la evaluación de la plataforma de vuelo , el diseño del cuadricóptero y del autopiloto, se han realizado distintos experimentos para intentar controlar la aeronave con algoritmos PID en cascada a una frecuencia de 70Hz.

Control en 1 GdL (*Pitch*)

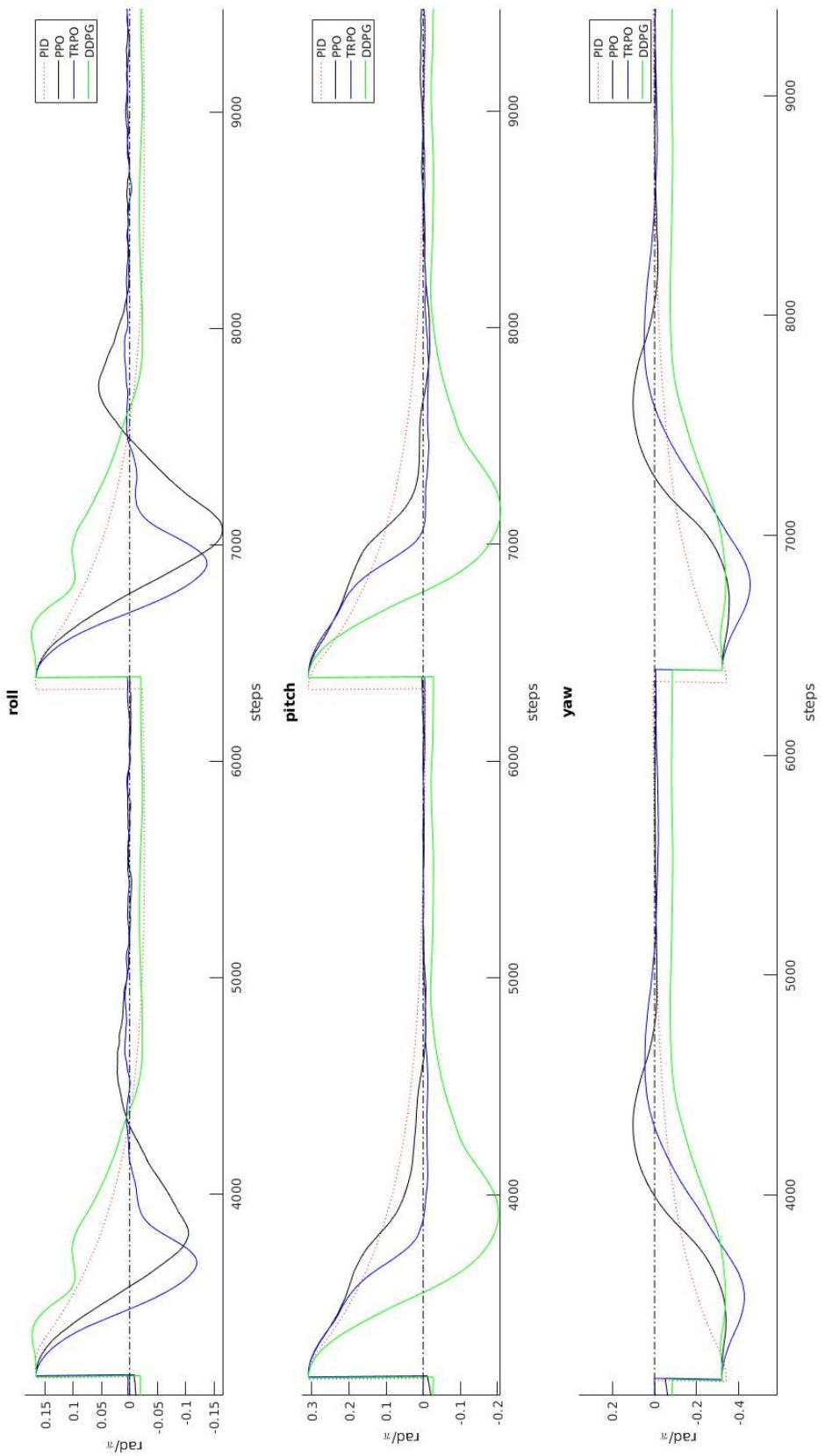
En este experimento (Fig. 7.4) se ha empleado la rótula de 1 GdL que únicamente permite el movimiento en *pitch*.

Se observa que el sistema es capaz de estabilizarse y que reacciona ante las perturbaciones externas de forma ágil. Por otro lado, también podemos observar ruido en la zona estable, además de poseer un pequeño error de posición de unos 2° .

Control en 1 GdL (*Roll*)

En este experimento (Fig. 7.6) se ha empleado la rótula de 1 GdL que únicamente permite el movimiento en *roll*.

El comportamiento observado es similar al del experimento anterior: el sistema se estabiliza y consigue llegar a la referencia después de perturbarlo de una forma ágil, pero también se observa ruido y un pequeño error de posición en el régimen permanente.

Figura 7.3: Estabilización en *roll*, *pitch* y *yaw* simultáneamente

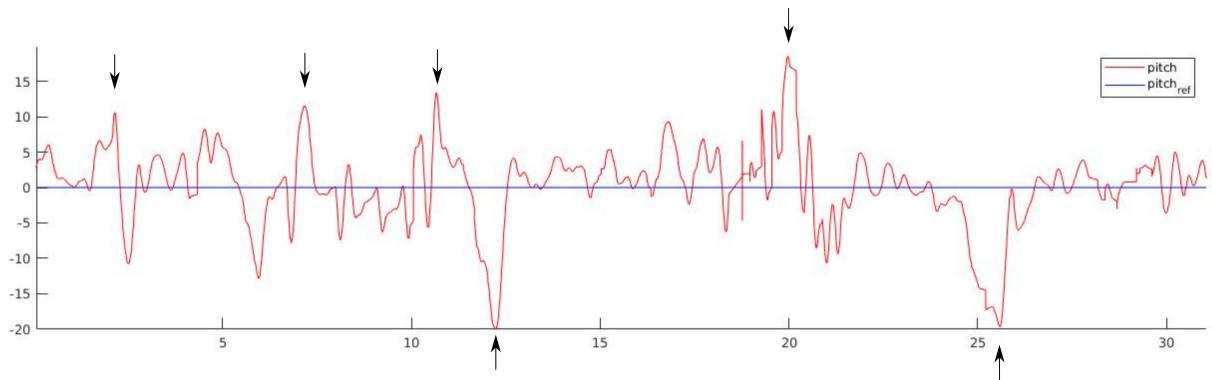


Figura 7.4: Estabilización en *pitch*. Las flechas marcan los tiempos en los que se perturbó al sistema.

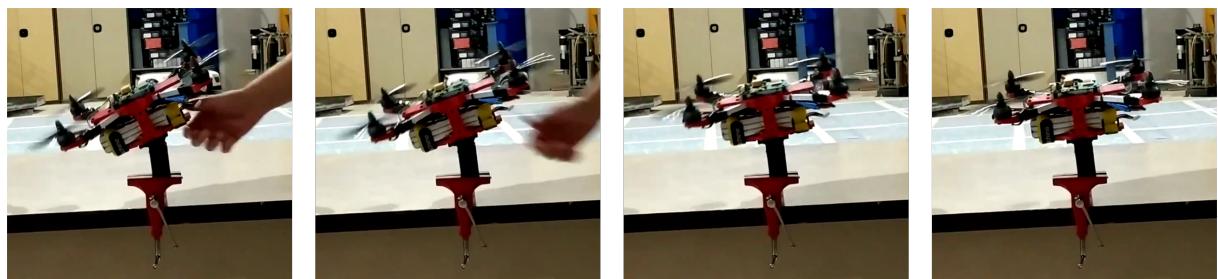


Figura 7.5: Estabilización en *pitch*

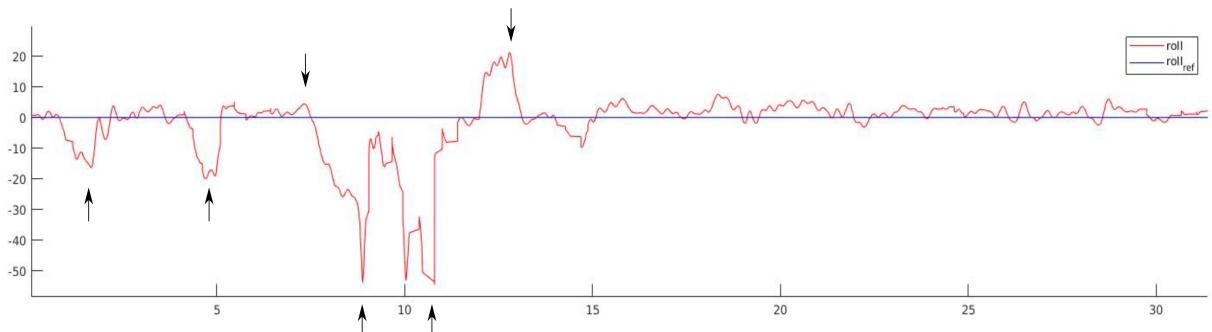


Figura 7.6: Estabilización en *roll*. Las flechas marcan los tiempos en los que se perturbó al sistema.



Figura 7.7: Estabilización en *roll*

Control en los 3 GdL

En este experimento se ha empleado la rótula de 1 GdL que permite el movimiento en *roll*, *pitch* y *yaw*.

En este experimento (Fig. 7.9) observamos que aumenta el ruido que encontramos en el régimen permanente en *pitch* y en *roll*, aunque se disminuye el error de posición que veíamos en los experimentos anteriores. Por otro lado el controlador de *yaw* es el menos ruidoso, pero también el más lento.

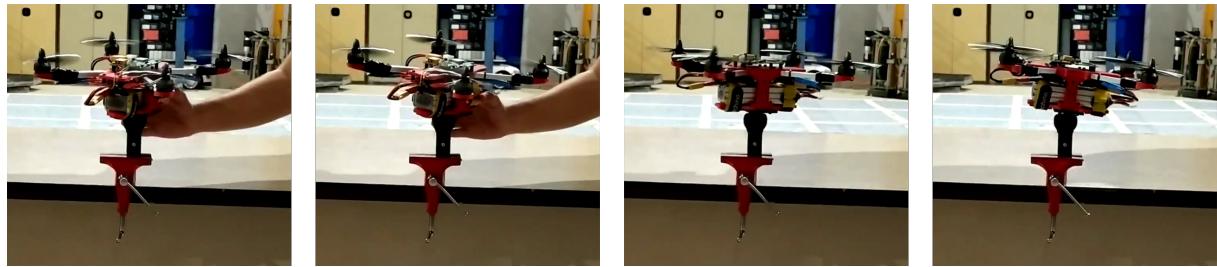


Figura 7.8: Estabilización en los 3 ejes

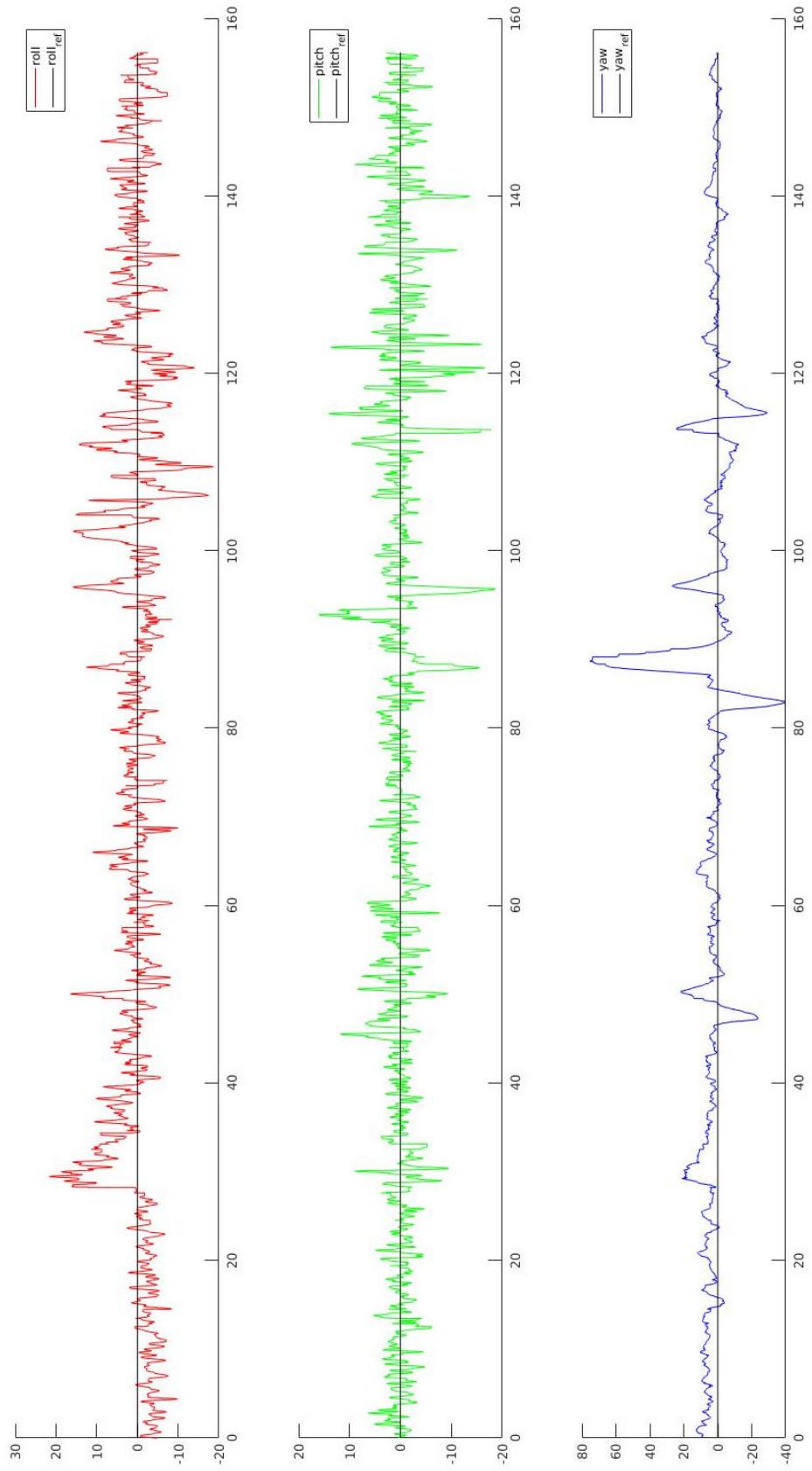


Figura 7.9: Estabilización en *roll*, *pitch* y *yaw* simultáneamente

Conclusiones y trabajo futuro

Durante el transcurso de este proyecto se ha conseguido desarrollar una plataforma, que permite estudiar los distintos algoritmos de control, con los que se puede estabilizar a un cuadricóptero. A continuación se hablará sobre las conclusiones que se han extraído durante este proceso y las posibles líneas de mejora que se podrían realizar en un futuro. En los apéndices se encuentra la planificación , el presupuesto y el impacto medioambiental del trabajo.

8.1. Conclusiones

Los objetivos principales que se plantearon al comienzo del proyecto consistían en: desarrollar un autopiloto propio capaz de ejecutar distintos algoritmos de control para cuadricópteros, desarrollar una plataforma de vuelo para poder probar este autopiloto y los algoritmos de forma segura e investigar sobre la posibilidad de emplear distintos algoritmos de aprendizaje por refuerzo para el control de actitud de la aeronave. A continuación desgranaremos las conclusiones que se han extraído de cada uno de estos objetivos y las posibles mejoras que se pueden realizar de cada uno.

Autopiloto

Se ha conseguido desarrollar un autopiloto funcional, capaz de estimar el estado de la aeronave y estabilizarse, cerrando un bucle interno de control y generando los comandos necesarios para controlar los motores. En este autopiloto se ha integrado la etapa de potencia necesaria para que sea posible alimentar al mismo directamente desde la batería, por lo que integra en una única PCB una gran cantidad de funcionalidades que no son habituales en los autopilotos comerciales, como por ejemplo la conexión WiFi, lo que ha permitido recabar los datos en tiempo real de forma sencilla. Sin embargo, las pistas que alimentan los motores a través de placa no se encuentran lo suficientemente ventiladas, por lo que se calientan demasiado.

Plataforma de vuelo

La plataforma de vuelo, ha permitido llevar a cabo los experimentos realizados de una forma segura y controlada. Debido a que la inmensa mayoría de las piezas han sido fabricadas mediante técnicas de impresión 3D, ha sido posible realizar variaciones de la plataforma de forma rápida y sencilla. Es por ésto que , se han diseñado varios componentes intercambiables, para poder modificar la configuración de los experimentos, lo que ha sido crucial para poder realizar el ajuste los parámetros del PID en las pruebas reales. El inconveniente que tiene la impresión 3D empleando PLA, es la fragilidad de algunas piezas, como por ejemplo, la unión esférica del banco de pruebas.

Algoritmos basados en aprendizaje por refuerzo

Uno de los retos que más tiempo han requerido, ha sido conseguir que los algoritmos propuestos, convergieran a una política óptima. Ha sido un proceso iterativo que ha tomado mucho tiempo hasta que se consiguió la convergencia del primer algoritmo. Posteriormente se realizaron muchas modificaciones de los hiperparámetros con el ánimo de mejorar el comportamiento del agente. Finalmente se ha conseguido aplicar varios algoritmos del estado del arte, consiguiendo , en algunos casos como en el del TRPO, muy buenos rendimientos.

8.2. Trabajo futuro

Esta plataforma, puede tener un gran interés para labores investigadoras y docentes. Para que pueda ser empleable de forma cómoda en estos ámbitos, sería posible simplificar el autopiloto, reduciendo así su tamaño y coste, ademas de reducir las dimensiones de la aeronave y el banco de pruebas, empleando por ejemplo motores DC, los cuales son mucho más pequeños y permiten controlarse de forma más sencilla. Si se reduce el tamaño, sería posible emplear varias de estas plataformas en los laboratorios de institutos y universidades para enseñar teoría de control. Además de reducir el tamaño, sería conveniente diseñar un interfaz más sencillo, pudiéndose integrar con programas como Matlab y Simulink, ésto mejoraría la usabilidad del mismo.

De cara a la investigación, sería conveniente realizar un modelo preciso de la aeronave en concreto, en vez de usar el modelo de otro cuadricóptero. Cuanto más parecido sea el modelo que se emplea en simulación al modelo real de la aeronave, menos se diferenciarán los comportamientos observados en simulación y el comportamiento real. En el desarrollo de algoritmos basados en aprendizaje por refuerzo, un aspecto muy relevante es el salto entre el entorno simulado y el entorno real, si el modelo real es lo suficientemente distinto al simulado, los comportamientos pueden variar enormemente. Es por esto que sería conveniente realizar una modelización dinámica completa de la aeronave.

En cuanto a los algoritmos, existen muchas maneras distintas de poder emplear el aprendizaje automático al campo del control de cuadricópteros, sería interesante replantearse la forma en la que los algoritmos actúan, para mejorar el rendimiento de estos algoritmos de forma que sobrepasen a los algoritmos de control clásico.Por ejemplo, en vez de generar un controlador en posición basado en RL, sería interesante diseñar un controlador en velocidad y sobre éste emplear un regulador P, basándose en la filosofía del controlador en cascada.

Esquemáticos del autopiloto

A continuación se adjuntan los esquemáticos del autopiloto diseñado.

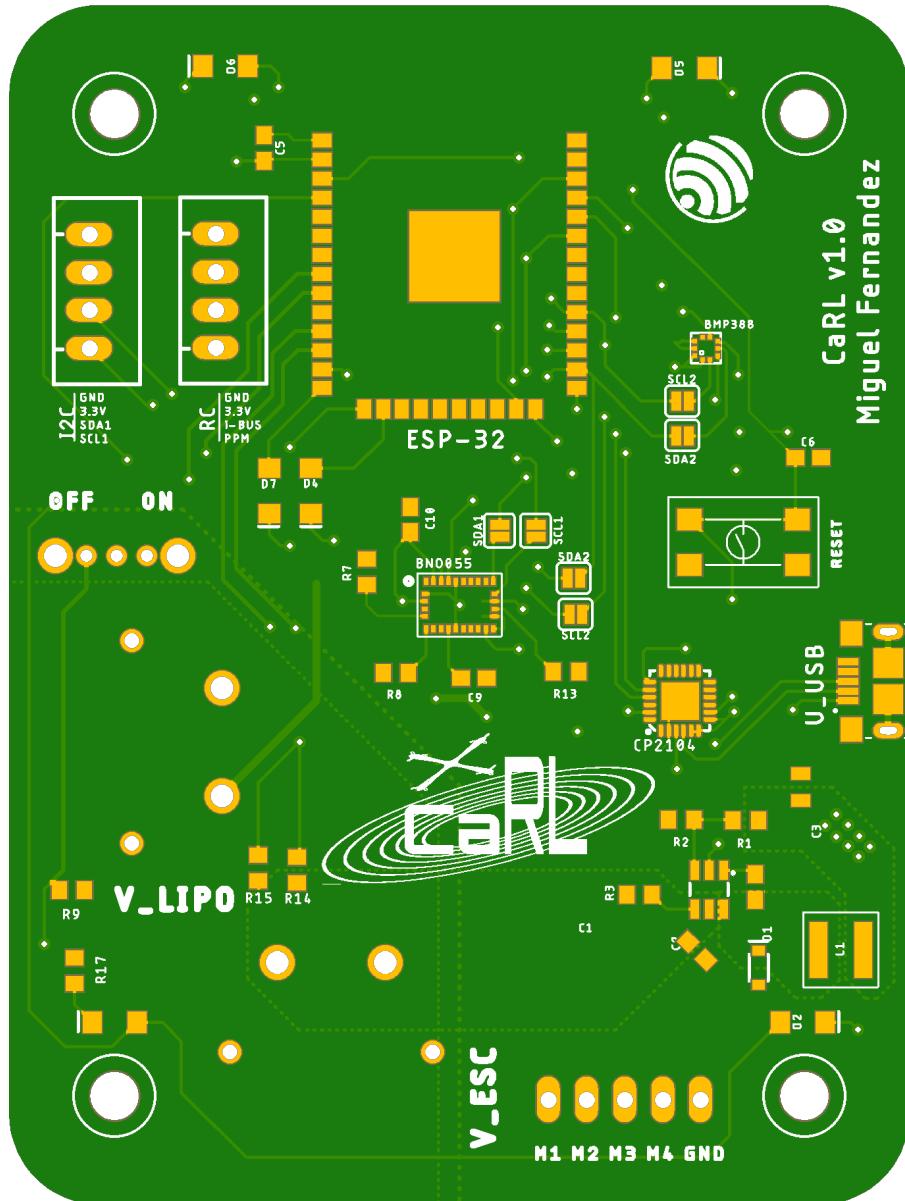
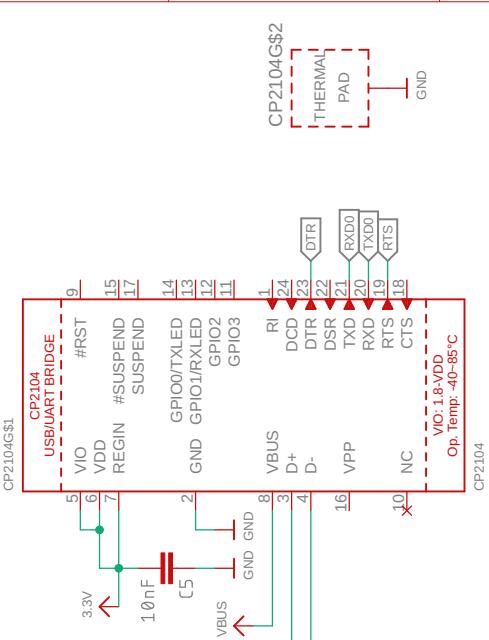


Figura A.1: *Board* PCB Autopiloto.

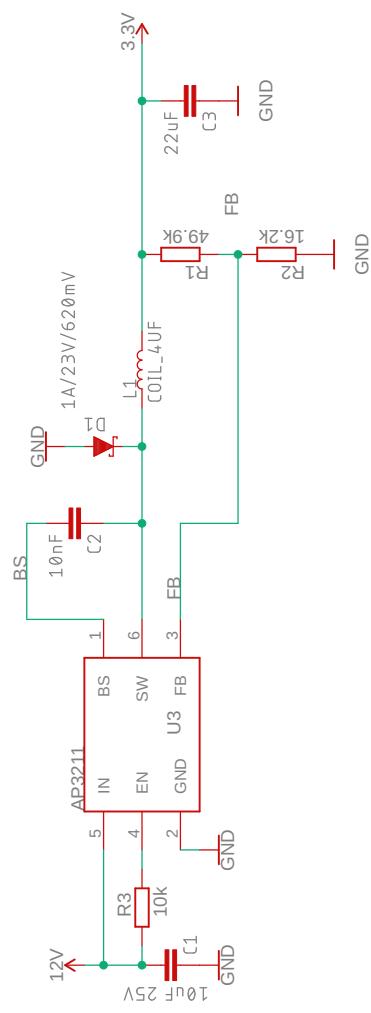
STEP-DOWN

1 2 3 4 5 6 7 8

USB TO SERIAL CONVERTER

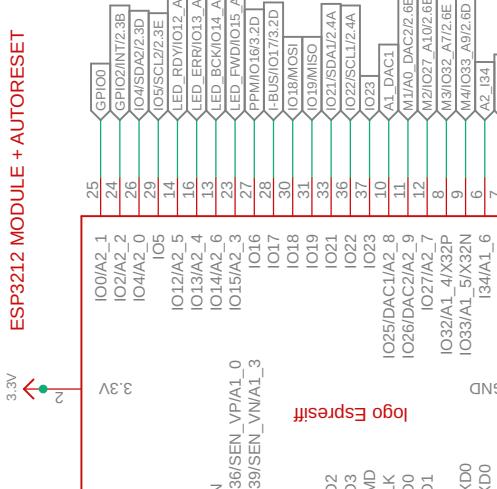


A

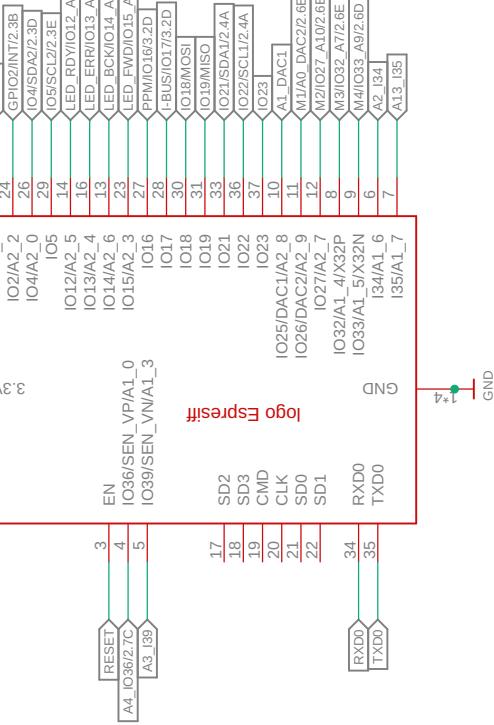


B

NC: IO0, IO2
VBAT SENSE: IO35
LED: IO13



C

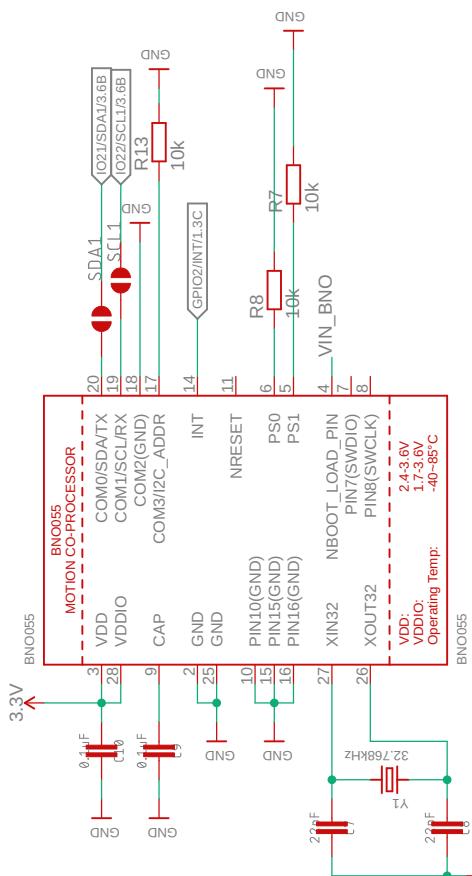


D

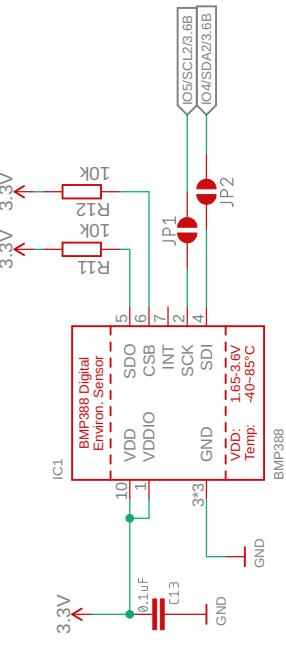


E

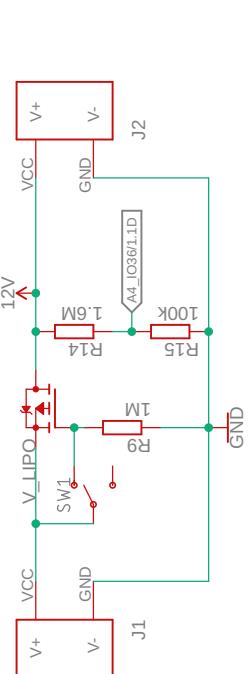
BNO055



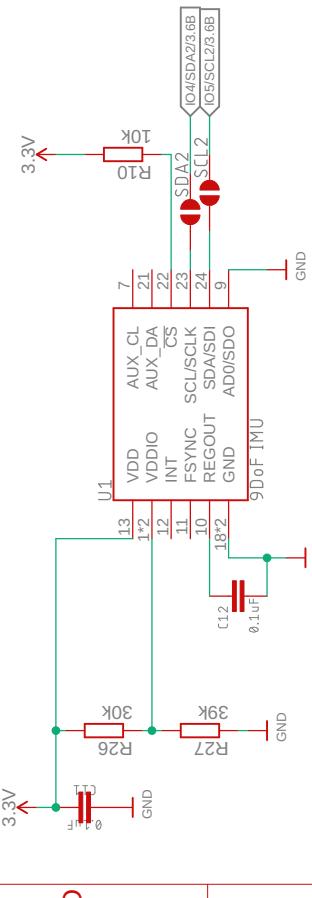
BMP388



POWER SWITCH



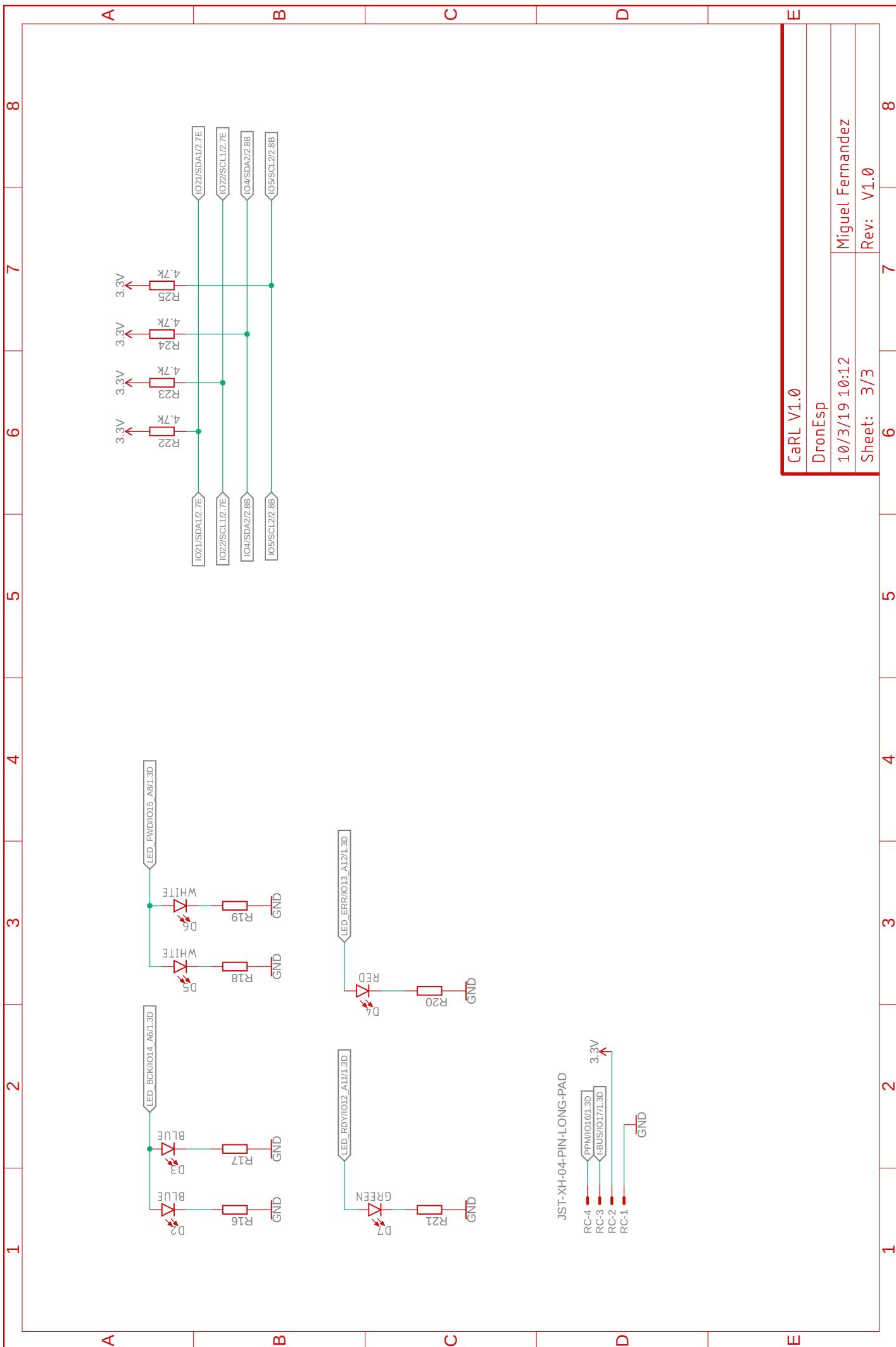
MPU9250



CONNECTORS



CaRL V1.0	DronEsp	10/3/19 10:12	Miguel Fernandez	Rev: V1.0
Sheet: 2/3				
6	7	8		



Presupuesto y Planificación

B.1. Presupuesto

El presupuesto del trabajo se puede separar en tres partes: recursos humanos, compra de material y amortización de los equipos utilizados.

En cuanto a los recursos humanos empleados, se ha tenido una dedicación por parte del alumno de unas 800 horas, esto es un número de horas mucho superior a las 360 horas (30h/ECTS) correspondientes a la carga temporal de los 12 ECTS del Trabajo fin de Grado (TFG). Esto se ha debido al gran alcance y a la complejidad del mismo. Un sueldo de investigador a media jornada en la universidad, sin estar graduado, es de unos 450 euros. Lo que se traduce en un salario de unos 5,625 euros la hora. Los salarios del tutor y el cotutor se han extraído del portal de transparencia de la UPM. La dedicación del tutor ha sido de unas 20 horas de implicación en el trabajo y la implicación del cotutor ha sido de unas 80 horas de implicación.

Recursos humanos	Horas	Coste Horario [EUR]	Total [EUR]
Alumno	800	5.625	4500
Cotutor	80	7.8	624
Tutor	20	33.72	674.4
Total			5798.4

Los costes de material del proyecto son debidos a la construcción del cuadricóptero y del autopiloto.

Material	Coste unitario [EUR]	Unidades	Total [EUR]
Cuadrcóptero			
Bobina PLA 1Kg	20	1.5	30
Perfiles aluminio	2	1	2
Pack 4 Motores MT2204 II	25	1	25
ESC BlHeli 4 in 1	50	1	50
Baterías LiPo	25	2	50
PCB autopiloto	20	1	20
Componentes PCB	50	1	50
Hélices HQ5040	2.5	4	10
Total			239

En cuanto a la amortización del equipo, se han empleado 2 ordenadores para el desarrollo del software y para el entrenamiento de los algoritmos. Se ha considerado una amortización lineal del 10 % de la vida útil (10 años).

Equipo	Precio	Coste Amortización(10 %)
Pc sobremesa	1980	198
Pc portátil	1300	130
Total		328

Añadiendo un coste de encuadernado de la memoria de unos 30 euros el presupuesto total del proyecto ha sido

Concepto	Total [EUR]
Recursos humanos	5798,4
Material	239
Amortización del equipo	328
Encuadernación	40
Total	6405,4

B.2. Planificación

La realización de este trabajo ha empleado un ritmo continuo de horas de trabajo desde su comienzo, siendo un poco menor en épocas de exámenes y un poco mayor al comienzo de los cuatrimestres y julio. La dedicación media diaria del trabajo ha sido de unas 4 horas semanales, durante un periodo de unos 10 meses (descontando agosto y septiembre), lo que da un total de unas 800 horas. La inmensa mayoría de estas horas se han dedicado en el Centro de Automática y Robótica (CAR) de la Escuela Técnica Superior de Ingenieros Industriales (ETSII) de la Universidad Politécnica de Madrid (UPM), concretamente en el grupo de investigación de Visión por Computador y Robots Aéreos (CVAR).

En cuanto a la distribución del trabajo en este tiempo, el trabajo comenzó a realizarse en septiembre de 2018, durante los primeros meses se realizó el curso sobre redes neuronales y aprendizaje profundo, en la plataforma online Coursera. La duración del curso se extendió hasta finales de diciembre. Paralelamente, a partir de octubre se comenzó con el diseño de la aeronave, y en noviembre con el del autopiloto. A principios de febrero se finalizó con el diseño y construcción del cuadricóptero y con el diseño y montaje de la PCB del autopiloto. A partir de este punto, el resto del tiempo se ha dedicado al software, tanto el del autopiloto, como el de la estación de tierra , al diseño de los algoritmos de control y a la experimentación real. Se ha realizado un diagrama GANTT (fig. B.1) en el que se ha detallado más en profundidad la distribución temporal de las tareas. Asimismo, se ha esquematizado la organización del proyecto en un diagrama EDP (fig. B.2).

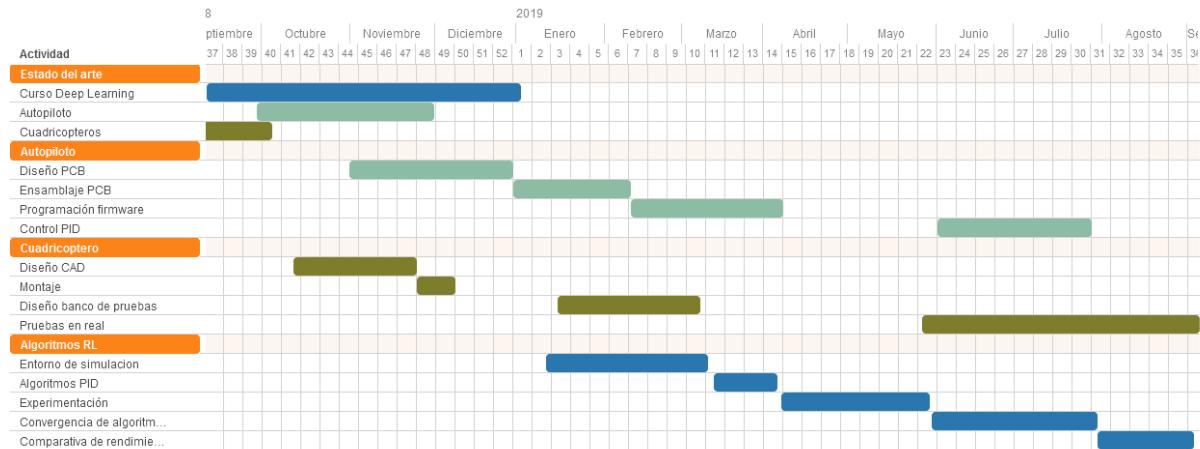


Figura B.1: Diagrama de Gantt

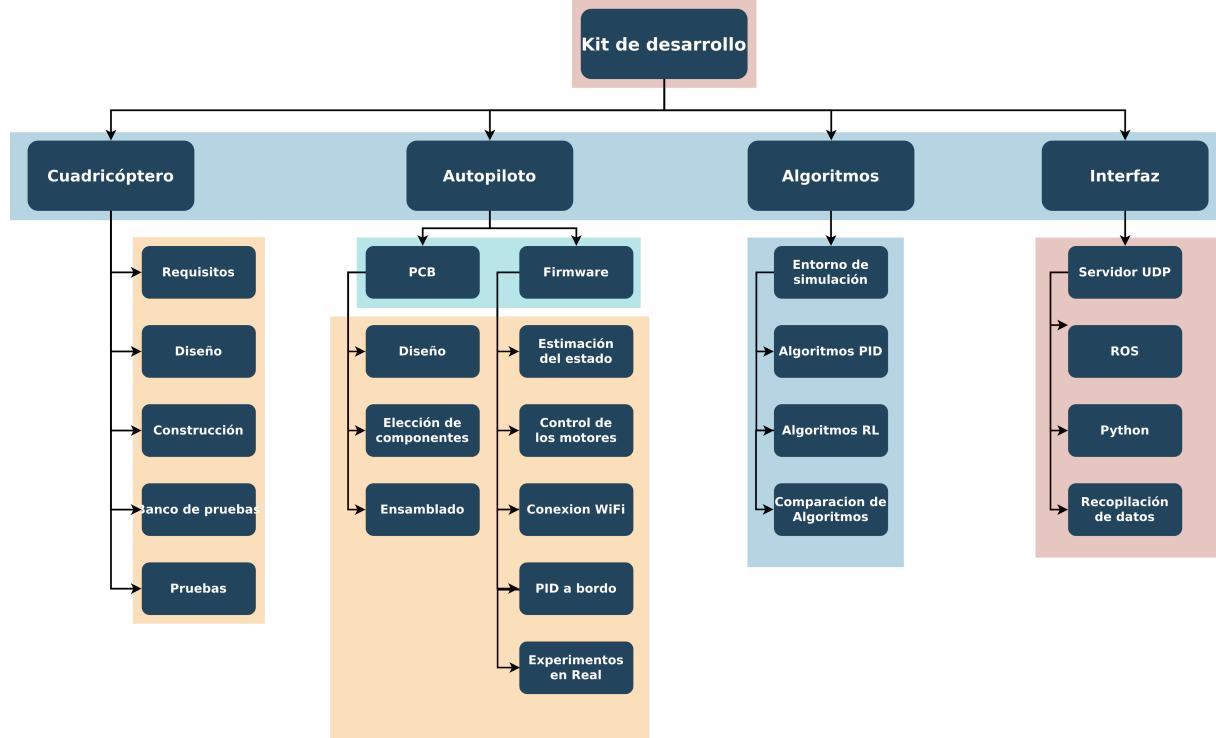


Figura B.2: Diagrama EDP

Impacto social y medioambiental

El impacto social que tiene este trabajo se ve reflejado en su posible empleo en la educación y la investigación. Actualmente las metodologías docentes están tendiendo hacia el aprendizaje práctico, hacia aprender haciendo. Esta plataforma podría emplearse en centros docentes debido a su montaje mecánico hecho casi en su totalidad con impresión 3D.

Desde el punto de vista de la investigación, tener la posibilidad de desarrollar y probar nuevos algoritmos para el control de cuadricópteros puede mejorar la efectividad del uso de estas aeronaves en múltiples aplicaciones. Cuanto mejor sea el controlador, más fácil será utilizar estas aeronaves para tareas de inspección, seguridad y búsqueda y rescate, entre otras.

El impacto medioambiental de la plataforma es reducido, ya que, el PLA es un plástico biodegradable y las baterías de Litio, una vez descargadas, son sencillas de desechar. EL proceso de fabricación de los componentes requiere de recursos materiales y energéticos, cuyo proceso de obtención puede provenir de fuentes no renovables. Sin embargo, los beneficios sociales que se pueden extraer de los resultados del proyecto hacen asumible este impacto medioambiental.

Índice de figuras

1	Autopiloto CaRL (<i>Cuadcopter with autopilot based on Reinforcement Learning</i>)	6
2	Cuadricóptero sobre el banco de pruebas	6
1.1	Esquema cuadricóptero en X	10
1.2	Autopiloto comercial Pixhawk 4	11
2.1	Esquema estructura gimbal	13
2.2	Plataforma FFT Gyro de Eureka Dynamics	14
2.3	Esquema de una union esférica	15
2.4	Plataforma 3 DOF Hover de la empresa Quanser	15
3.1	Bucle de control realimentado con regulador PID	17
3.2	Bucle de control en cascada	18
3.3	Esquema cuadrirrotor en X	18
3.4	Esquema de una red neuronal artificial	19
3.5	Esquema de un perceptrón	20
3.6	Función sigmoide	20
3.7	Función tangente hiperbólica	20
3.8	Función ReLu	21
3.9	Diagrama canónico del bucle de interacción entorno-agente	22
3.10	Evolución de una función límite inferior con un algoritmo MM.	27
3.11	Gráficas de la función objetivo $L^{\text{CLIP}}(\theta)$ para distintos valores de r , dependiendo del signo de la ventaja A	27
4.1	Cuadricóptero diseñado con el autopiloto incorporado.	28
4.2	Cuadricóptero diseñado con el autopiloto incorporado.	29
4.3	Portamotores en CAD	29
4.4	Núcleo en CAD	30
4.5	Separadores en CAD	30
4.6	Portabaterías en CAD	31
4.7	Motores LHI MT2204 II empleados	31
4.8	Tabla de especificaciones motor MT2204 II.	32
4.9	Hélices tripala HQ5040	32
4.10	Funcionamiento ESC (www.hwtomechatronics.com)	33
4.11	ESC Multistar Race 4 in 1 30A BLHeli empleado	33
4.12	Batería LiPo 3s 35C 5200 mAh de la marca NZACE empleada	34
4.13	PCB autopiloto CaRL, anverso y reverso.	34
4.14	Esquema de un convertidor Buck	35
4.15	Sensores BNO055 y BMP388 respectivamente	37
4.16	Rotulas de 1GdL (<i>pitch</i> y <i>roll</i> respectivamente)	37

4.17	Junta esférica antes de ensamblar	38
4.18	Junta esférica doble	38
5.1	Entorno de simulación GymFC en Gazebo 9	41
5.2	Esquema interfaz Estacion-Autopiloto	43
6.1	Formas de onda que recibe un variador	45
6.2	Funciones R_t para distintos valores de $n = 2, 4, 10$ respectivamente	46
7.1	Estabilización en <i>roll</i> en un entorno simulado.	47
7.2	Estabilización en <i>pitch</i> en un entorno simulado.	48
7.3	Estabilización en <i>roll</i> , <i>pitch</i> y <i>yaw</i> simultáneamente	49
7.4	Estabilización en <i>pitch</i> . Las flechas marcan los tiempos en los que se perturbó al sistema.	50
7.5	Estabilización en <i>pitch</i>	50
7.6	Estabilización en <i>roll</i> . Las flechas marcan los tiempos en los que se perturbó al sistema.	50
7.7	Estabilización en <i>roll</i>	50
7.8	Estabilización en los 3 ejes	51
7.9	Estabilización en <i>roll</i> , <i>pitch</i> y <i>yaw</i> simultáneamente	52
A.1	<i>Board</i> PCB Autopiloto.	55
B.1	Diagrama de Gantt	61
B.2	Diagrama EDP	61

Bibliografía

- [1] M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray *et al.*, “Learning dexterous in-hand manipulation,” *arXiv preprint arXiv:1808.00177*, 2018.
- [2] H. J. Kim, M. I. Jordan, S. Sastry, and A. Y. Ng, “Autonomous helicopter flight via reinforcement learning,” in *Advances in neural information processing systems*, 2004, pp. 799–806.
- [3] A. Y. Ng and M. Jordan, “Pegasus: A policy search method for large mdps and pomdps,” in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 406–415.
- [4] A. Y. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang, “Autonomous inverted helicopter flight via reinforcement learning,” in *Experimental robotics IX*. Springer, 2006, pp. 363–372.
- [5] T. Dierks and S. Jagannathan, “Output feedback control of a quadrotor uav using neural networks,” *IEEE transactions on neural networks*, vol. 21, no. 1, pp. 50–66, 2010.
- [6] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter, “Control of a quadrotor with reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2096–2103, 2017.
- [7] W. Koch, R. Mancuso, R. West, and A. Bestavros, “Reinforcement learning for uav attitude control,” *ACM Transactions on Cyber-Physical Systems*, vol. 3, no. 2, p. 22, 2019.
- [8] W. Koch, R. Mancuso, and A. Bestavros, “Neuroflight: Next generation flight control firmware,” *arXiv preprint arXiv:1901.06553*, 2019.
- [9] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, p. 529, 2015.
- [11] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.

- [12] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” 2014.
- [13] G. E. Uhlenbeck and L. S. Ornstein, “On the theory of the brownian motion,” *Physical review*, vol. 36, no. 5, p. 823, 1930.
- [14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [15] “Esp32 technical reference manual,” Espressif Systems, 2018. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32_technical_reference_manual_en.pdf
- [16] “Esp32 datasheet,” Espressif Systems, 2019. [Online]. Available: https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf
- [17] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “Openai gym,” *arXiv preprint arXiv:1606.01540*, 2016.
- [18] A. Hill, A. Raffin, M. Ernestus, A. Gleave, R. Traore, P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, and Y. Wu, “Stable baselines,” <https://github.com/hill-a/stable-baselines>, 2018.
- [19] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, Y. Wu, and P. Zhokhov, “Openai baselines,” <https://github.com/openai/baselines>, 2017.