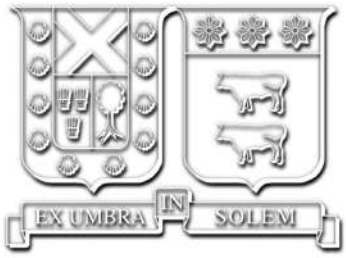
Universidad Técnica Federico Santa María
Departamento de Ingeniería Eléctrica
Valparaíso, Chile



Análisis y diseño de control basado en inteligencia artificial para aplicaciones de convertidores de potencia monofásicos y trifásicos

Matías Javier Kaemling Arévalo

2026

Requisito parcial para obtener el título de:
Ingeniero Civil Electricista

Profesora Guía:
Dr. Margarita Norambuena Valdivia (UTFSM)
Profesor Correferente:
Dr. Alejandro Angulo Cárdenas (UTFSM)

Valparaíso, marzo, 2026.

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ■ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Análisis y diseño de control basado en inteligencia artificial para aplicaciones de convertidores de potencia monofásicos y trifásicos.

Nombre del candidato(a): Matías Javier Kaemling Arévalo

Carrera / Grado: Ingeniería Civil Eléctrica

Campus: Casa Central **Departamento:** Departamento de Ingeniería Eléctrica

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Margarita Norambuena Valdivia, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

■ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por **(marcar una opción)**:

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

_____

_____

_____

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

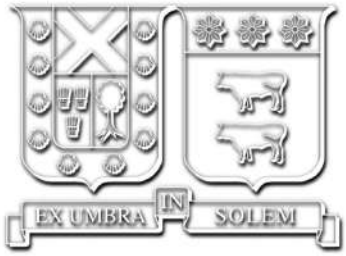
Fecha: 31/03/2026 **Firma:** _____

Estudiante o Candidato(a):

Fecha:31/03/2026 **Firma:** _____

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Universidad Técnica Federico Santa María
Departamento de Ingeniería Eléctrica
Valparaíso, Chile



Análisis y diseño de control basado en inteligencia artificial para aplicaciones de convertidores de potencia monofásicos y trifásicos

Matías Javier Kaemling Arévalo

2026

Agradecimientos

En primer lugar, quiero agradecer a mi madre, quien me dio la vida y ha sido un apoyo incondicional en cada etapa de este camino. A pesar de todas las dificultades, siempre estuvo a mi lado, acompañándome y luchando para que fuera un niño feliz.

Al amor de mi vida, América, con quien hemos formado nuestra pequeña familia perruna, gracias por ser mi sostén en los momentos más difíciles y por acompañarme siempre con amor y paciencia. Gracias también a Chokita y Bluey, quienes nos llenan de alegría y amor incondicional cada día.

Quiero también agradecer a Ronald, a quien agradezco de corazón por haber elegido ser mi padre. Gracias a él conocí el mundo de la electricidad, despertando un interés que terminaría marcando mi camino profesional.

Agradezco profundamente a mi segunda familia: Milenko, Olinda y Jorge, quienes me acogieron con cariño durante la pandemia y me hicieron sentir como uno más de la familia. Su generosidad, cercanía y constante apoyo han sido un pilar muy importante en mi vida.

Quiero agradecer a mis amigos de la universidad, quienes hicieron de este proceso algo más llevadero, acompañándome en los desafíos, aprendizajes y momentos difíciles. Su apoyo y compañerismo fueron fundamentales a lo largo de este camino.

Finalmente, agradezco de manera muy especial a la profesora Margarita, cuyo apoyo, guía y disposición fueron clave para poder sacar esto adelante. Sin su ayuda, este logro no habría sido posible.

Índice de Contenidos

Índice de Contenidos I

Índice de Figuras VII

Índice de Tablas XI

Índice de Algoritmos XIII

Nomenclatura 1

1 Introducción 1

1.1 Motivación . . . 1

1.2 Objetivo general . . . 2

1.3 Objetivos específicos . . . 2

2 Marco Teórico 3

2.1 Convertidores de potencia e inversores de dos niveles . . . 4

2.2 Modelado de cargas . . . 4

2.3 Estrategias clásicas de control . . . 6

2.3.1 Control proporcional integral (PI) de corriente . . . 6

2.3.2 Control de campo orientado (FOC) de MAS . . . 7

2.4 Fundamentos del aprendizaje por refuerzo . . . 8

2.5 Algoritmos basados en aprendizaje por refuerzo . . . 10

2.5.1 Q-Learning . . . 11

2.5.2 Deep Q-Network (DQN) . . . 13

3 Configuración del Sistema y de los Agentes de Aprendizaje 19

3.1 Modelo eléctrico del sistema . . . 20

3.1.1 Sistema monofásico . . . 20

3.1.2 Sistemas trifásicos . . . 20
3.2 Formulación del Proceso de Decisión de Markov . . . 21
3.2.1 Conjunto de estados . . . 21
3.2.2 Conjunto de acciones . . . 22
3.2.3 Función de recompensa . . . 22
3.2.4 Episodios de aprendizaje . . . 23
3.2.5 Consideraciones adicionales . . . 23
3.3 Control de corriente monofásico con agente Q-learning . . . 24
3.3.1 Plataforma de simulación . . . 24
3.3.2 Representación del agente Q-learning en la simulación . . . 25
3.3.3 Selección de parámetros de Q-learning . . . 26
3.3.4 Diseño del experimento de evaluación de parámetros . . . 27
3.4 Control de corriente monofásico con agente DQN . . . 28
3.4.1 Plataforma de entrenamiento . . . 28
3.4.2 Representación del agente DQN en la simulación . . . 29
3.4.2.a Espacio de observaciones y acciones . . . 29
3.4.2.b Configuración de la Q-Network . . . 29
3.4.2.c Opciones de entrenamiento del agente . . . 30
3.4.3 Selección de parámetros de Deep Q-Network . . . 30
3.4.4 Diseño del experimento de evaluación de parámetros . . . 31
3.4.4.a Fase I: Exploración Amplia del Espacio de Hiperparámetros . . . 32
3.4.4.b Fase II: Ajustes de la Dinámica del Entorno y Estabilidad del Aprendizaje 33
3.4.4.c Fase III: Análisis Univariable y Evaluación de Funciones de Recompensa 33
3.5 Configuración de los agentes trifásicos de control de corriente . . . 34
3.5.1 Agente Q-learning trifásico en marco $\alpha\beta$ . . . 34
3.5.2 Agente Q-learning trifásico en marco dq . . . 34
3.5.3 Agente DQN trifásico en marco $\alpha\beta$ . . . 35
3.5.4 Agente DQN trifásico en marco dq . . . 35

4 Implementación de esquemas de control en PLECS 36

4.1 Implementación del agente Q learning en *PLECS Standalone* . . . 37
4.2 Implementación del Controlador DQN en PLECS mediante Tabla LUT . . . 38
4.3 Implementación de los controles trifásicos de corriente en PLECS . . . 39
4.3.1 Estructura general del control trifásico . . . 39

4.3.2 Implementación del control trifásico Q-learning en PLECS 39
4.3.3 Implementación del control trifásico DQN en PLECS 41
4.3.4 Diferencias prácticas entre Q-learning y DQN en PLECS 42

5 Análisis de impacto de hiperparámetros en aprendizaje de agentes 43

5.1 Métricas de desempeño 44
5.2 Control de corriente monofásico con agente Q-learning 45
5.2.1 Análisis univariado de los hiperparámetros 45
5.2.1.a Efecto de la tasa de aprendizaje 45
5.2.1.b Efecto del factor de descuento 46
5.2.1.c Efecto de la exploración 46
5.2.1.d Efecto del paso de discretización 47
5.2.1.e Conclusiones 48
5.2.2 Mapa de calor parámetros 49
5.2.3 Ranking de combinaciones evaluadas 49
5.3 Control de corriente monofásico con agente DQN 52
5.3.1 Fase I: Exploración Amplia del espacio de Hiperparámetros 52
5.3.2 Fase II: Ajuste de la dinámica del entorno y estabilidad de aprendizaje . . 55
5.3.3 Fase III: Análisis Univariado y Evaluación de Funciones de Recompensa . 56
5.3.3.a Tasa de aprendizaje 56
5.3.3.b Factor de descuento 57
5.3.3.c Número de neuronas por capa 57
5.3.3.d Tamaño del minibatch 58
5.3.3.e Decaimiento del parámetro de exploración 58
5.3.3.f Factor de suavizado del objetivo (*Target Smooth Factor*) 59
5.3.3.g Función de recompensa 59
5.3.4 Entrenamiento Final 60

6 Análisis de controles de corrientes implementados en PLECS 63

6.1 Control de corriente monofásico con agente Q-learning 64
6.1.1 Control de corriente frente a cambio de resistencia 64
6.1.2 Control de corriente frente a cambio de inductancia 64
6.1.3 Control de corriente frente a cambio frecuencia 66
6.1.4 Control de corriente frente a cambio de amplitud 67
6.1.5 Control de corriente frente a referencia constante 68

6.1.6 Efecto de Action Hold en control de señales cuadradas 69
6.1.7 Ajustes alternativos para el control Q-learning 69
6.1.8 Evaluación de mecanismos estabilizadores de control 71
6.1.9 Respuesta de control frente a distintas función de recompensas 73
6.1.10 Impacto de la dimensión de tabla Q 75
6.2 Control de corriente monofásico con agente Deep Q-Network 77
6.2.1 Control de corriente frente a cambio de resistencia 77
6.2.2 Control de corriente frente a cambio de inductancia 78
6.2.3 Control de corriente frente a cambio de frecuencia 79
6.2.4 Control de corriente frente a cambio de amplitud 80
6.2.5 Seguimiento de señal cuadrada 81
6.2.6 Impacto de construcción de LUT 81
6.3 Control de corriente trifásico con agente Q-learning marco $\alpha\beta$ 83
6.3.1 Control de corriente frente a diferentes amplitudes 84
6.3.2 Control de corriente frente a variaciones de resistencia de carga 85
6.3.3 Control de corriente frente a variaciones de inductancia de carga 86
6.3.4 Control de corriente frente a variaciones de frecuencia eléctrica 87
6.4 Control de corriente trifásico con agente Q-learning marco *dq* 88
6.4.1 Control de corriente frente a diferentes amplitudes 89
6.4.2 Control de corriente frente a variaciones de resistencia de carga 90
6.4.3 Control de corriente frente a variaciones de inductancia de carga 91
6.4.4 Control de corriente frente a variaciones de frecuencia eléctrica 92
6.5 Control de corriente trifásico con agente DQN en marco $\alpha\beta$ 93
6.5.1 Control de corriente frente a diferentes amplitudes 93
6.5.2 Control de corriente frente a variaciones de resistencia de carga 94
6.5.3 Control de corriente frente a variaciones de inductancia de carga 95
6.5.4 Control de corriente frente a variaciones de frecuencia eléctrica 95
6.6 Control de corriente trifásico con agente DQN marco *dq* 96
6.6.1 Control de corriente frente a diferentes amplitudes 97
6.6.2 Control de corriente frente a variaciones de resistencia de carga 97
6.6.3 Control de corriente frente a variaciones de inductancia de carga 98
6.6.4 Control de corriente frente a variaciones de frecuencia eléctrica 99

7 Control de velocidad en máquina asincrónica 101

7.1 Control de velocidad QL-FOC para motor de inducción . . . 102
7.1.1 Arranque de MAS QL-FOC . . . 102
7.1.2 Seguimiento de referencia de velocidad QL-FOC . . . 103
7.1.3 Cambio de carga mecánica QL-FOC . . . 105
7.2 Control de velocidad DQN-FOC para motor de inducción . . . 107
7.2.1 Arranque de MAS DQN-FOC . . . 107
7.2.2 Seguimiento de referencia de velocidad DQN-FOC . . . 108
7.2.3 Cambio de carga mecánica DQN-FOC . . . 110
7.3 Comparación de controles con IFOC . . . 111
7.3.1 Tensión aplicada por inversor . . . 111
7.3.2 Corrientes inyectadas por inversor . . . 113
7.3.3 Torque electromagnético producido sobre el motor . . . 114
7.3.4 Comparación de métricas de distintos métodos . . . 114
7.4 Comentarios finales de control de velocidad . . . 116

8 Conclusiones 117

APÉNDICE 118

A Mecanismos adicionales de estabilización de la política de acción 119

B Resultados complementarios de entrenamiento en control de corriente 120
B.1 Control de corriente monofásico con agente Q-learning . . . 120
B.1.1 Evaluación de mejores y peores casos de agentes según RMSE . . . 120

C Fundamentos matemáticos y límites físicos 123
C.1 Diferencias finitas . . . 123
C.2 Límite de corriente de inversor monofásico con carga RL . . . 123

D Medidas estadísticas representativas en experimentos de aprendizaje por refuerzo 124
D.1 Mediana . . . 124

E Motor de inducción trifásico 126
E.1 Parámetros de motor . . . 126
E.2 Parámetros de controladores . . . 127
E.2.1 PI del lazo de velocidad . . . 127

E.2.2 FOC tradicional: lazo interno de corriente 127

F Resultados adicionales de control de velocidad 128

Referencias 131

Índice de Figuras

2.1 Esquema eléctrico de inversor de dos niveles monofásico conectado a carga RL. 4

2.2 Esquema de control de campo orientado indirecto (IFOC). 8

2.3 Interacción agente-entorno en un proceso de decisión de Markov [12]. 9

2.4 Esquema de Q-learning como MDP. 12

2.5 Esquema genérico de una red neuronal artificial con N neuronas en cada capa oculta. 14

2.6 Esquema de Deep Q-Network como MDP. 15

3.1 Esquema eléctrico de inversor de dos niveles trifásico conectado a carga trifásica. 20

3.2 Esquema general de la plataforma de simulación para el caso monofásico con agente Q-learning. 24

3.3 Esquema general de la plataforma de entrenamiento para el caso monofásico con agente DQN. 29

4.1 Control de corriente con agente Q-learning en PLECS Standalone. 37

4.2 Control de corriente con agente Deep Q-Network en PLECS Standalone. 38

4.3 Control de corriente trifásico en marco $\alpha\beta$ con agente Q-learning en PLECS Standalone. 40

4.4 Control de corriente trifásico en marco dq con agente Q-learning en PLECS Standalone. 40

5.1 Mediana RMSE para cada valor de la tasa de aprendizaje α. 46

5.2 Mediana RMSE para cada valor del factor de descuento γ. 46

5.3 RMSE promedio para cada valor de la probabilidad de exploración ε. 47

5.4 Mediana RMSE para cada valor del paso de discretización e_{step}. 47

5.5 RMSE promedio del seguimiento de corriente en función de α y γ. 49

5.6 Control de corriente con el agente Q-learning de mejor desempeño (Run 381). . 50

5.7 Control de corriente con un agente de desempeño intermedio (Run 346). 51

5.8 Control de corriente con el agente Q-learning de peor desempeño (Run 55). . . 51

5.9 Recompensa promedio de entrenamientos de agentes en primer barrido de parámetros. 53
5.10 Evaluación de agente DQN con mejor desempeño (RMSE) en primer barrido. . 54
5.11 Recompensa promedio para entrenamiento de agentes DQN en 1000 episodios. 55
5.12 Evaluación de agente DQN con mejor desempeño (RMSE) en barrido de mil episodios. 56
5.13 Desempeño de agentes DQN para tasa de aprendizaje α. 57
5.14 Desempeño de agentes DQN para el factor de descuento γ. 57
5.15 Desempeño de agentes DQN para distintos números de neuronas N_n por capa. 58
5.16 Desempeño de agentes DQN para el tamaño del minibatch $|\mathcal{M}|$. 58
5.17 Desempeño de agentes DQN para el decaimiento del parámetro $\varepsilon_{\text{decay}}$. 59
5.18 Desempeño de agentes DQN para el parámetro *Target Smooth Factor*. 59
5.19 Desempeño de agentes DQN para distintas funciones de recompensa. 60
5.20 Evaluación de agente DQN con mejor desempeño (RMSE) en barrido final. . . . 61

6.1 Respuesta del control Q-learning ante variaciones de la resistencia de carga en PLECS. 64
6.2 Respuesta del control Q-learning ante variaciones de la inductancia de carga en PLECS. 65
6.3 Respuesta cambio de frecuencia con $\text{A}_{\text{ref}} = 10$ A - agente QL en PLECS. 66
6.4 Respuesta cambio de amplitud a $f_{\text{ref}} = 50$ Hz - agente QL en PLECS. 67
6.5 Respuesta a una señal escalón - agente QL en PLECS. 68
6.6 Efecto de parámetro Action Hold en la respuesta a señal cuadrada - agente QL en PLECS. 69
6.7 Efecto de parámetro Action Hold en la respuesta a cambio de inductancia. . . . 70
6.8 Efecto del parámetro Action Hold ante un cambio de referencia sinusoidal a cuadrada. 71
6.9 Efecto de parámetros de control frente a un cambio de referencia sinusoidal a cuadrada 72
6.10 Respuesta de control de corriente con agente Q-learning para distintas recompensas. 74
6.11 Control de corriente con agente Q-learning con Dim(Q) = 483. 75
6.12 Control de corriente con agente Q-learning con Dim(Q) = 123. 76
6.13 Respuesta cambio de resistencia de carga - agente DQN en PLECS. 77
6.14 Respuesta cambio de inductancia de carga - agente DQN en PLECS. 78
6.15 Respuesta cambio de frecuencia con $\text{A}_{\text{ref}} = 10$ A - agente DQN en PLECS. . . . 79
6.16 Respuesta ante cambios de amplitud con $f_{\text{ref}} = 50$ Hz — agente DQN en PLECS. 80

6.17 Seguimiento de referencia cuadrada - agente DQN en PLECS. 81
6.18 Seguimiento de referencia con LUT de diferente rango - agente DQN en PLECS. 81
6.19 Control de corriente en marco de referencia $\alpha\beta$ con agente Q-learning. 83
6.20 Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la amplitud de referencia. 84
6.21 Respuesta del control de corriente en marco de referencia $\alpha\beta$ ante cambios en la resistencia. 85
6.22 Respuesta del control de corriente en marco de referencia $\alpha\beta$ ante cambios en la inductancia. 86
6.23 Desempeño del control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la frecuencia eléctrica. 87
6.24 Control de corriente en marco de referencia dq con agente Q-learning. 88
6.25 Control de corriente con agente Q-learning en marco de referencia dq. 89
6.26 Control de corriente con agente Q-learning en marco de referencia dq. 90
6.27 Control de corriente en marco de referencia dq con agente Q-learning. 91
6.28 Control de corriente en marco de referencia dq con agente Q-learning. 92
6.29 Control de corriente en marco de referencia $\alpha\beta$. 93
6.30 Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la amplitud de referencia. 94
6.31 Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la resistencia de la carga. 94
6.32 Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la inductancia de la carga. 95
6.33 Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la frecuencia de referencia. 96
6.34 Control de corriente en marco de referencia dq. 96
6.35 Control de corriente en marco de referencia dq frente a cambios en la amplitud de referencia. 97
6.36 Control de corriente en marco de referencia dq frente a cambios en la resistencia de la carga. 98
6.37 Control de corriente en marco de referencia dq frente a cambios en la inductancia de la carga. 99
6.38 Control de corriente en marco de referencia dq frente a cambios en la frecuencia de referencia. 99

7.1 Esquema de control propuesto QL-IFOC y DQN-IFOC para motor de inducción trifásico. 101
7.2 Arranque de MAS a velocidad de referencia $\omega_{\text{ref}} = 0.4 \cdot \omega_{\text{nom}}$ con QL-FOC. 103

7.3 Cambio de velocidad de referencia sin carga mecánica con QL-FOC. 104
7.4 Cambio de carga mecánica a velocidad constante con QL-FOC. 105
7.5 Arranque de MAS a velocidad nominal sin carga con DQN-FOC. 108
7.6 Cambio de velocidad de referencia sin carga mecánica con DQN-FOC. 109
7.7 Cambio de carga mecánica a velocidad constante con DQN-FOC. 110
7.8 Comparación en tensión aplicada en estado estacionario sin carga entre FOC y DQN-FOC. 111
7.9 Comparación en tensión aplicada en estado estacionario sin carga entre FOC y QL-FOC. 112
7.10 Comparación en corriente inyectada por inversor en estado estacionario sin carga entre FOC y DQN-FOC. 113
7.11 Comparación en torque electromagnético generado por el motor en estado estacionario sin carga entre FOC y DQN-FOC. 114

F.1 Arranque de MAS a velocidad de referencia $\omega_{\text{ref}} = \omega_{\text{nom}}$ con QL-FOC. 128
F.2 Arranque de MAS a velocidad de referencia $\omega_{\text{ref}} = \omega_{\text{nom}}$ y flujo debilitado con QL-FOC. 129

Índice de Tablas

3.1 Vectores de conmutación de un inversor trifásico de dos niveles 22

3.2 Parámetros explorados en Q-learning 27

3.3 Parámetros considerados en la exploración de DQN. 31

5.1 Resumen de estadísticas de RMSE para cada hiperparámetro del algoritmo Q-learning. 48

5.2 Los 10 mejores y 10 peores resultados del entrenamiento Q-learning según RMSE. 50

5.3 Parámetros considerados en la exploración del primer barrido DQN. 52

5.4 Resultados del entrenamiento DQN según RMSE para primer barrido de parámetros. 54

5.5 Hiperparámetros de entrenamiento final. 61

5.6 Resultados entrenamiento final. 62

6.1 Métricas obtenidas para distintas variaciones de inductancia en el control Q-learning. 65

6.2 Métricas del barrido en frecuencia para el control Q-learning. 66

6.3 Métricas para distintas amplitudes de corriente en el control Q-learning. 67

6.4 Comparación de distorsión para cambio de inductancia $f_1 = 50$ Hz. 70

6.5 Resumen de métricas por configuración para los tramos de 10 mH y 5 mH, incluyendo el espacio para registrar el valor RMS de la corriente medida bajo referencia cuadrada. 72

6.6 Métricas en estado estacionario para diferentes funciones de recompensa. . . . 74

6.7 Hiperparámetros considerados para la exportación de la LUT. 77

6.8 Métricas obtenidas para distintas variaciones de inductancia en el control Deep Q-Network. 78

6.9 Métricas del barrido en frecuencia para el control Deep Q-Network. 80

6.10 Métricas para distintas amplitudes de corriente en el control Deep Q-Network. 80

7.1 Comparación de métricas en operación sin carga. 115

7.2 Comparación de métricas en operación con carga nominal. 115

B.1 Resultados del entrenamiento DQN en 1000 episodios. 120
B.2 Resultados del entrenamiento DQN según RMSE con distintas funciones de recompensas R. 121
B.3 Resultados del entrenamiento DQN según RMSE con distintos valores de TSF. . 121
B.4 Resultados entrenamiento final. 122

E.1 Parámetros eléctricos y mecánicos de la máquina de inducción trifásica. 126
E.2 Valores nominales de la máquina y corrientes en el marco dq. 126
E.3 Parámetros del PI del lazo de velocidad. 127
E.4 Parámetros del lazo interno de corriente del FOC tradicional. 127

Índice de Algoritmos

2.1 Q-learning para control en un entorno genérico [12] 13
2.2 Deep Q-Network con *experience replay* 17
3.1 Interacción agente-entorno mediante funciones *Reset* y *Step* 32

Nomenclatura

α	Tasa de aprendizaje
γ	Factor de descuento
$\hat{Q}(s, a; \theta^-)$	Red objetivo (*target Q-network*) con parámetros θ^-.
$\mathcal{A}$	Conjunto de acciones del agente
$\mathcal{L}_{\mathrm{DQN}}(\theta)$	Función de pérdida del DQN (pérdida cuadrática temporal).
$\mathcal{S}$	Conjunto de estados
MDP	Proceso de decisión de Markov (*Markov Decision Process*).
ω	Velocidad mecánica del rotor
ω_{ref}	Referencia de velocidad mecánica
τ	*Target smooth factor* (TSF), factor de suavizado para actualización blanda de la red objetivo.
θ	Parámetros (pesos) de la *Q-network*.
θ^-	Parámetros (pesos) de la red objetivo (*target network*).
θ_e	Ángulo eléctrico del sistema
θ_{mec}	Ángulo mecánico del rotor
ε	Parámetro de exploración en política ε-greedy.
ε	Parámetro de exploración
$\varepsilon_{\mathrm{decay}}$	Tasa de decaimiento de ε
$\varepsilon_{\mathrm{decay}}$	Tasa de decaimiento de exploración.
e_α, e_β	Errores de corriente en el marco $\alpha\beta$
e_d, e_q	Errores de corriente en el marco dq
e_N	Error normalizado (definido según la métrica empleada en el texto).
e_{step}	Paso de discretización del error de corriente (Q-learning).
$e_{\mathrm{máx}}$	Límite superior del error de corriente para discretización (Q-learning).

$e_{\text{mín}}$ Límite inferior del error de corriente para discretización (Q-learning).

f_s Frecuencia de muestreo, $f_s = 1/T_s$.

$f_{\text{disc}}(e(t))$ Función de discretización que asigna el error continuo $e(t)$ a un estado discreto.

f_{pwm} Frecuencia de modulación PWM

f_{sw} Frecuencia efectiva de conmutación

i_α, i_β Componentes de corriente en el marco de referencia estacionario $\alpha\beta$

i_a, i_b, i_c Corrientes de fase del sistema trifásico

i_d, i_q Componentes de corriente en el marco de referencia rotante dq

$i_d^{\text{ref}}, i_q^{\text{ref}}$ Referencias de corriente en el marco dq

J Inercia mecánica del conjunto rotor–carga

K_p, K_i Ganancias proporcional e integral del controlador PI

L_m Inductancia magnetizante

L_s, L_r Inductancias totales del estator y del rotor

N_θ Número de sectores angulares

N_e Número de estados discretos asociados a la discretización del error.

N_n Número de neuronas por capa en la red neuronal.

N_{meas} Número de niveles discretos considerados para la corriente medida.

N_{ref} Número de niveles discretos considerados para la corriente de referencia.

p Número de pares de polos de la máquina

$Q(s, a)$ Función de valor acción–estado

$Q(s, a; \theta)$ Función de valor-acción aproximada por red neuronal (*Q-network*), parametrizada por θ.

R_r' Resistencia del rotor referida al estator

R_s Resistencia del estator

T_e Torque electromagnético

T_L Torque mecánico de carga

T_s Período de muestreo.

v_a, v_b, v_c Tensiones de fase del inversor trifásico

V_{dc} Tensión de fuente continua

DQN *Deep Q-Network*.

DQN–FOC Control de velocidad tipo IFOC con lazo interno de corriente reemplazado por una política DQN (LUT).

FOC Control de campo orientado (*Field-Oriented Control*).

IFOC Control de campo orientado indirecto (*Indirect Field-Oriented Control*).

LUT *Look-Up Table*

MAS Máquina Asincrónica

PI Proporcional-Integral

PWM Pulse Width Modulation / Modulación por ancho de pulso

QL Q-learning.

QL–FOC Control de velocidad tipo IFOC con lazo interno de corriente reemplazado por Q-learning.

RMSE Error cuadrático medio

SV Space Vector

THD Distorsión armónica total

TSF *Target smooth factor.*

Capítulo 1

Introducción

1.1. Motivación

En últimas décadas, los convertidores de potencia se han convertido en un pilar fundamental en la estructura eléctrica moderna. Dada su capacidad de transformar, regular y controlar el flujo de energía los hace imprescindibles en una amplia gama de aplicaciones, que incluye sistemas de generación basado en energías renovables, accionamientos eléctricos industriales, movilidad eléctrica, sistemas de almacenamiento energético, entre muchos otros. En este contexto, los inversores de potencia desempeñan un rol clave por tener la capacidad de convertir energía entre corriente continua y corriente alterna, buscando una integración eficiente entre fuentes de energía, sistemas de potencia y consumos finales.

El desempeño de estos sistemas depende en gran medida de la precisión y robustez de la estrategia de control implementada, la cual determina la calidad de la forma de onda entregada y la estabilidad del sistema frente a variaciones de carga o perturbaciones. En sus primeras implementaciones, los convertidores fueron controlados mediante estrategias clásicas como el control proporcional integral derivativo (*PID*), ampliamente utilizado gracias a su simplicidad, bajo costo y robustez en sistemas lineales [1]. Si bien estos métodos han demostrado eficacia en condiciones nominales, su desempeño se ve limitado ante la presencia de no linealidades, incertidumbre paramétrica o variaciones rápidas del entorno, lo que motiva la búsqueda de enfoques de control más adaptativos e inteligentes.

A medida que han transcurrido las décadas, los procesos industriales se han vuelto cada vez más complejos debido a la incertidumbre inherente a sus dinámicas, lo que ha impulsado el desarrollo de técnicas de control más avanzadas, como el control adaptativo y el control predictivo basado en modelos (*Model Predictive Control*), capaces de anticipar y ajustar el comportamiento del sistema en función de su evolución temporal y de las variaciones en sus condiciones de operación [2].

A pesar de los avances alcanzados con estrategias como el control adaptativo y el control predictivo basado en modelos, la mayoría de los métodos tradicionales mantienen una fuerte dependencia del modelo matemático del sistema. Históricamente, los controles empleados en inversores de potencia se han basado en esquemas lineales o modulares, tales como el proporcional–integral (PI), el control por histéresis o el control deadbeat, los cuales ofrecen un desempeño adecuado en condiciones nominales y sistemas bien caracterizados [3, 4]. No obstante, su efectividad disminuye cuando el sistema enfrenta perturbaciones externas, variaciones de carga o no linealidades significativas, ya que requieren un modelado preciso y una sintonización fina de parámetros. Estas limitaciones han

impulsado la exploración de enfoques de control más flexibles y autónomos, capaces de adaptarse al entorno sin depender de un modelo explícito del sistema [5, 6].

En los últimos años, el avance de la inteligencia artificial ha abierto nuevas posibilidades para el control de sistemas dinámicos, destacando el aprendizaje por refuerzo (*Reinforcement Learning, RL*) como una estrategia capaz de "aprender" políticas de control óptimas directamente a partir de la interacción con el entorno, sin requerir un modelo exacto del sistema y con la capacidad de adaptarse a condiciones operativas variables [7].

En este contexto, la incorporación de la inteligencia artificial en áreas cómo la electrónica de potencia, donde el control de convertidores e inversores debe enfrentarse a variaciones de carga, sistemas no lineales y restricciones de eficiencia. La combinación de algoritmos de aprendizaje por refuerzo con entornos de simulación avanzados como MATLAB/simulink y PLECS ofrece una plataforma segura para desarrollar y evaluar controladores inteligentes antes de su implementación real. En este marco, el presente trabajo estudia la aplicación de Q-leaning (*QL*) y Deep Q-Network (DQN) al control de corriente en diferentes tipos de cargas, incluyendo carga RL trifásica, una máquina de inducción, buscado demostrar el potencial de estas técnicas de control en sistemas eléctricos.

1.2. Objetivo general

Analizar y diseñar controles basados Q-learning para aplicaciones de convertidores de potencia en casos monofásicos y trifásicos mediante simulación en software PLECS.

1.3. Objetivos específicos

I. Analizar control base Q-learning en circuito monofásico con carga estática y para circuitos trifásicos con carga RL y motor de inducción.

II. Analizar control Q-learning adaptativo con redes neuronales (Deep Q Network) en circuitos monofásicos y trifásicos para cargas RL y motor de inducción trifásico.

III. Diseñar control Q-learning y DQN (Deep Q Network) para implementarse en un entorno de simulación empleando el software PLECS para el entorno de potencia y Matlab para el diseño y ajuste de los controladores.

IV. Comparar el desempeño de los controladores diseñados y realizar un análisis de los mismos.

Capítulo 2

Marco Teórico

A lo largo de este capítulo se presentan las bases conceptuales y técnicas que sustentan el desarrollo de este trabajo. Se exponen brevemente los principios de funcionamiento de los convertidores de potencia, con especial énfasis en el inversor de dos niveles, así como las características de las cargas consideradas en el estudio: una carga resistiva–inductiva (RL) y un motor asincrónico.

Luego, se realiza un breve repaso de las estrategias de control clásicas empleadas en la regulación de corriente y velocidad, particularmente el control proporcional–integral (*PI*) y el control de campo orientado (*FOC*), utilizados como referencia para evaluar el desempeño de los métodos propuestos en este trabajo.

Finalmente, se presentan los fundamentos del aprendizaje por refuerzo, profundizando en los algoritmos basados en Q-Learning, y destacando su relevancia en el control de sistemas no lineales y su potencial para optimizar el desempeño de convertidores de potencia sin requerir un modelo matemático exacto.

2.1. Convertidores de potencia e inversores de dos niveles

Los convertidores de potencia constituyen un elemento esencial en sistemas modernos de conversión y control de energía, presente en aplicaciones de accionamientos eléctricos, energías renovables, electromovilidad, entre otros. Estos dispositivos electrónicos permiten transformar la energía eléctrica, permitiendo realizar modificaciones de tensión, corriente y frecuencia según se necesite [8].

Los inversores de potencia corresponden a un subgrupo de convertidores y se encargan de transformar la energía de corriente continua en corriente alterna, regulando amplitud, frecuencia y forma de onda de la tensión o corriente de salida. Entre las topologías más compactas existentes se encuentra el inversor de dos niveles, el cual destaca por su simplicidad y bajo coste de implementación. Este convertidor emplea transistores bipolares de puerta aislada (IGBTs, por sus siglas en inglés) junto con diodos en antiparalelo, que permiten la conducción bidireccional de la corriente y generación de tensiones alternas [9].

La configuración de dos niveles es la base para el desarrollo de topologías más complejas, como inversores multinivel o modulación vectorial, por lo que representa un buen punto de partida para el análisis de las estrategias de control desarrolladas en el presente trabajo.

2.2. Modelado de cargas

El sistema de estudio se compone de un inversor de dos niveles, alimentado por una fuente de corriente continua (DC), encargado de generar las tensiones que alimentan la carga monofásica o trifásica. En la Figura 2.1 se muestra el esquema eléctrico correspondiente al caso monofásico analizado, donde S_1 y S_2 representan las señales de conmutación que controlan la conducción de los IGBT respectivos, mientras que $\overline{S_1}$ y $\overline{S_2}$ corresponden a sus negaciones lógicas. Para el caso trifásico, se emplea un inversor de dos niveles análogo, utilizado tanto con carga estática como con el motor de inducción.

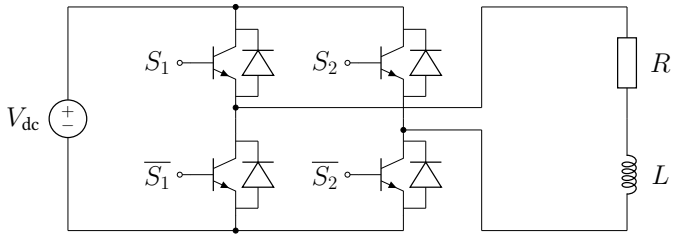


Figura 2.1: Esquema eléctrico de inversor de dos niveles monofásico conectado a carga RL.

Para caso monofásico la Ecuación 2.1 es la que rige el sistema, dado que se trata de lazo cerrado en donde la tensión aplicada corresponde al valor entregado en la salida del inversor.

$$L\frac{di}{dt} + Ri = v \tag{2.1}$$

donde L es la inductancia, R la resistencia de la carga, i la corriente instantánea y v la tensión aplicada por el inversor.

Para la carga trifásica estática, se extiende el modelo presentado para el caso monofásico. La carga se conecta en configuración estrella (Y), y al tratarse de un sistema equilibrado, las tensiones de fase

corresponden a las tensiones entre fase y neutro. Las ecuaciones que describen el comportamiento de cada fase son equivalentes y se expresan como:

$$L\frac{di_p}{dt} + Ri_p = v_{pn} \tag{2.2}$$

donde i_p corresponde a la corriente instantánea de fase y v_{pn} a las tensión instantánea entre la fase $p = \{a, b, c\}$, y el punto neutro.

La representación de las corrientes trifásicas en sistemas de coordenadas ortogonales simplifica el análisis y el diseño de estrategias de control, al permitir desacoplar las componentes de magnitud y fase. para ello, se aplica primero la Transformación de Clarke, que proyecta las variables trifásicas (a, b, c) sobre un plano estacionario (α, β), y posteriormente la Transformación de Park, que rota este plano un ángulo θ_s para obtener las componentes en el sistema síncrono (d, q). De este modo, las variables del sistema pueden expresarse en un marco de referencia giratorio vinculado al campo magnético, lo que facilita la implementación de técnicas de control vectorial como el *Field Oriented Control (FOC)* [3]. Las transformaciones se expresan de la siguiente forma:

$$\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix}\begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \tag{2.3}$$

donde i_α e i_β corresponden a las corrientes en el plano estacionario (α, β) obtenidas a partir de las corrientes de fase i_a, i_b e i_c.

$$\begin{bmatrix} i_d \\ i_q \end{bmatrix} = \begin{bmatrix} \cos\theta_s & \sin\theta_s \\ -\sin\theta_s & \cos\theta_s \end{bmatrix}\begin{bmatrix} i_\alpha \\ i_\beta \end{bmatrix} \tag{2.4}$$

donde i_d e i_q corresponden a las corrientes en el plano sincrónico (d, q) obtenidas a partir de las corrientes i_α e i_β.

$$\begin{bmatrix} i_d \\ i_q \end{bmatrix} = \frac{2}{3}\begin{bmatrix} \cos\theta_s & \cos\left(\theta_s - \frac{2\pi}{3}\right) & \cos\left(\theta_s + \frac{2\pi}{3}\right) \\ -\sin\theta_s & -\sin\left(\theta_e - \frac{2\pi}{3}\right) & -\sin\left(\theta_s + \frac{2\pi}{3}\right) \end{bmatrix}\begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \tag{2.5}$$

donde i_d e i_q corresponden a las corrientes en el plano sincrónico (d, q) obtenidas a partir de las corrientes de fase i_a, i_b e i_c.

Para el análisis y control de la máquina de inducción en este trabajo, se emplea un modelo dinámico formulado en el sistema de referencia sincrónico (d, q), con todas las magnitudes referidas al estator. Las corrientes y tensiones del estator se expresan como componentes directas d y en cuadratura q, lo que permite representar en forma aproximada desacoplada las dinámicas asociadas al flujo y al torque electromagnético. Este enfoque es particularmente adecuado para la implementación de estrategias de *Field Oriented Control (FOC)*, donde se regulan las corrientes i_d e i_q de manera independiente.

El modelo considera las ecuaciones de tensión del estator, los enlaces de flujo y la dinámica mecánica del eje, y se expresa en las ecuaciones 2.6-2.9 [10, 9]:

$$\begin{aligned} v_{ds} &= R_s i_{ds} + \frac{d\psi_{ds}}{dt} - \omega_s \psi_{qs}, \\ v_{qs} &= R_s i_{qs} + \frac{d\psi_{qs}}{dt} + \omega_s \psi_{ds}, \end{aligned} \tag{2.6}$$

$$\begin{aligned} \psi_{ds} &= L_s i_{ds} + L_m i_{dr}, \quad \psi_{qs} = L_s i_{qs} + L_m i_{qr} \\ \psi_{dr} &= L_r i_{dr} + L_m i_{ds}, \quad \psi_{qr} = L_r i_{qr} + L_m i_{qs} \end{aligned} \tag{2.7}$$

$$T_e = \frac{3}{2} p L_m \left(i_{qs} i_{dr} - i_{ds} i_{qr} \right) \tag{2.8}$$

$$J \frac{d\omega_r}{dt} = T_e - T_L - B\omega_r \tag{2.9}$$

donde v_{ds} y v_{qs} son las tensiones del estator en el sistema de referencia síncrono, i_{ds} e i_{qs} las corrientes del estator, ψ_{ds} y ψ_{qs} los enlaces de flujo del estator, ψ_{dr} y ψ_{qr} los enlaces de flujo del rotor referidos al estator, R_s la resistencia del estator, L_s y L_r las inductancias propias del estator y rotor, L_m la inductancia mutua, ω_s la velocidad síncrona, ω_r la velocidad angular del rotor, p el número de pares de polos, J la inercia del conjunto rotórico, B el coeficiente de fricción viscosa, T_e el torque electromagnético y T_L el torque de carga.

En el contexto del control FOC, el sistema de referencia (d, q) se alinea comúnmente con el flujo del rotor, de modo que la componente i_d se asocia principalmente al control del flujo y i_q al control del torque electromagnético [9].

2.3. Estrategias clásicas de control

2.3.1. Control proporcional integral (PI) de corriente

El control proporcional integral (PI) es una de las estrategias más empleadas en sistemas de conversión y control de potencia debido a su simplicidad, bajo costo computacional y buen desempeño en sistemas lineales. Su objetivo es minimizar el error entre la señal de referencia y la variable controlada mediante la acción combinada de un término proporcional, que corrige el error instantáneo, y un término integral, que elimina el error en estado estacionario. Su ley de control se expresa como:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau \tag{2.10}$$

donde $e(t) = r(t) - y(t)$ es el error entre la referencia $r(t)$ y la salida $y(t)$, mientras que K_p y K_i representan las ganancias proporcional e integral, respectivamente.

En el contexto del control de corriente en convertidores, el controlador PI genera una tensión de referencia que, aplicada al inversor, permite regular la corriente de carga siguiendo una forma de onda deseada (por ejemplo, sinusoidal o cuadrada). Pese a su efectividad en condiciones nominales, su desempeño se ve afectado por variaciones en los parámetros del sistema o perturbaciones externas, lo que limita su adaptabilidad [4].

2.3.2. Control de campo orientado (FOC) de MAS

El control de campo orientado (*Field Oriented Control*, FOC) es una técnica de control vectorial que permite regular de forma independiente el flujo y el torque electromagnético de una máquina asincrónica (MAS), logrando un comportamiento dinámico similar al de una máquina de corriente continua. Para ello, las variables trifásicas del estator se transforman al sistema de referencia sincrónico (d, q), donde en estado estacionario las componentes de corriente pueden aproximarse como constantes, lo que facilita el uso de controladores lineales como el control PI [9, 10].

En el esquema de FOC con orientación al flujo del rotor, el eje d del sistema de referencia síncrono se alinea con el vector de flujo del rotor ψ_r. Bajo esta condición, la componente de flujo en el eje en cuadratura es nula ($\psi_{qr} \approx 0$), y la dinámica de la máquina se desacopla, de forma que la corriente i_{ds} se asocia al control del flujo, mientras que la corriente i_{qs} determina el torque electromagnético. De esta forma, la expresión del torque puede escribirse como [10, 11]:

$$T_e = \frac{3}{2}\, p\, \frac{L_m}{L_r}\, \psi_r\, i_{qs}, \tag{2.11}$$

donde p es el número de pares de polos, L_m la inductancia mutua, L_r la inductancia del rotor, ψ_r el flujo del rotor e i_{qs} la componente de corriente del estator en el eje q.

Bajo este enfoque, el control del flujo se realiza especificando una referencia ψ_r^* (o, de forma equivalente, una referencia para i_{ds}^*), mientras que el control del torque se lleva a cabo imponiendo una referencia de torque T_e^*, la cual se traduce en una referencia de corriente i_{qs}^*. A partir de (2.11), la corriente de referencia en el eje q puede aproximarse como:

$$i_{qs}^* = \frac{2}{3}\, \frac{L_r}{p\, L_m}\, \frac{T_e^*}{\psi_r^*}. \tag{2.12}$$

En la implementación clásica del FOC, el lazo interno de corrientes se encarga de regular las componentes i_{ds} e i_{qs} hacia sus respectivas referencias i_{ds}^* e i_{qs}^*. Definiendo los errores de corriente como $e_d(t) = i_{ds}^*(t) - i_{ds}(t)$ y $e_q(t) = i_{qs}^*(t) - i_{qs}(t)$, las tensiones de referencia en el plano sincrónico pueden obtenerse mediante controladores PI independientes:

$$v_{ds}^*(t) = K_{pd}\, e_d(t) + K_{id} \int_0^t e_d(\tau)\, d\tau, \tag{2.13}$$

$$v_{qs}^*(t) = K_{pq}\, e_q(t) + K_{iq} \int_0^t e_q(\tau)\, d\tau, \tag{2.14}$$

donde K_{pd}, K_{id}, K_{pq} y K_{iq} son las ganancias proporcional e integral de los lazos de corriente en los ejes d y q, respectivamente. Estas tensiones de referencia (v_{ds}^*, v_{qs}^*) se transforman de vuelta al sistema de coordenadas estacionario (α, β) y posteriormente al sistema trifásico (a, b, c) mediante las transformaciones inversas de Park y Clarke, generando las referencias de tensión para el inversor que alimenta la máquina.

El FOC proporciona una respuesta rápida y un control preciso del torque y del flujo, por lo que se ha convertido en un estándar industrial para el accionamiento de motores de inducción. Sin embargo, su implementación práctica requiere conocer o estimar el ángulo de orientación del flujo del rotor, lo que puede realizarse mediante distintas estrategias.

Una de las implementaciones más utilizadas es el control de campo orientado indirecto (*Indirect Field Oriented Control*, IFOC), en el cual el ángulo del sistema de referencia sincrónico se obtiene a partir de la velocidad angular eléctrica de la máquina, calculada como la suma de la velocidad mecánica del rotor y el deslizamiento. De esta forma, el ángulo sincrónico puede determinarse integrando la velocidad eléctrica, evitando la necesidad de medir directamente el flujo del rotor [10].

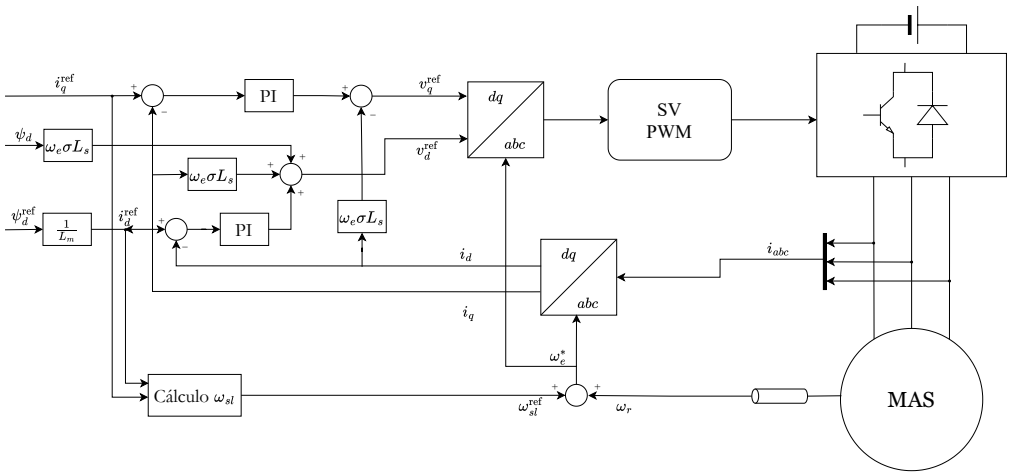


Figura 2.2: Esquema de control de campo orientado indirecto (IFOC).

La Figura 2.2 muestra el esquema general del control de campo orientado indirecto utilizado en este trabajo. En este enfoque, el lazo externo de velocidad genera la referencia de corriente i_{qs}^*, mientras que la referencia de flujo se fija mediante i_{ds}^*. A partir de las corrientes medidas y de la velocidad mecánica del rotor, se estima el flujo del rotor y se calcula la frecuencia de deslizamiento, lo que permite obtener el ángulo síncrono necesario para las transformaciones de Park y Clarke utilizadas en el control vectorial.

2.4. Fundamentos del aprendizaje por refuerzo

Antes de abordar las técnicas de aprendizaje por refuerzo, es necesario comprender el marco teórico en el que se fundamentan. Este se describe mediante los *Procesos de Decisión de Markov Finitos* (*Finite Markov Decision Process*, MDP). Como se señala en el Capítulo 3 del libro *Reinforcement Learning: An Introduction* de Sutton y Barto [12], un MDP es un modelo matemático que permite representar problemas de toma de decisiones secuenciales bajo incertidumbre, en los cuales un *agente* interactúa con un *entorno* en una serie de pasos discretos.

El *agente* corresponde a una entidad o elemento que aprende y toma decisiones, mientras que todo lo externo a él constituye el *entorno*. Ambos interactúan de manera secuencial: el agente selecciona una acción en función del estado actual, y el entorno responde a dicha acción generando un nuevo estado y una recompensa asociada. El objetivo del agente es maximizar, a lo largo del tiempo, la suma acumulada de recompensas recibidas mediante la elección de acciones adecuadas.

En la Figura 2.3 se muestra un esquema que ilustra esta interacción entre el agente y el entorno. En cada paso de tiempo t, el *agente* recibe una representación del *estado* actual del entorno, $S_t \in \mathcal{S}$, y en función de este selecciona una *acción*, $A_t \in \mathcal{A}(S_t)$. Un instante de tiempo después, como consecuencia de dicha acción, el agente obtiene una *recompensa* $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$ y el entorno

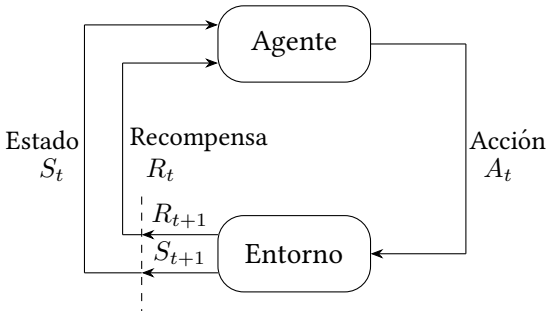


Figura 2.3: Interacción agente-entorno en un proceso de decisión de Markov [12].

transita a un nuevo *estado*, S_{t+1}. Los conjuntos de *estados* $\mathcal{S}$, *acciones* $\mathcal{A}$ y *recompensas* $\mathcal{R}$ son finitos y definen completamente el espacio de interacción del agente dentro del proceso de decisión de Markov [12]. Formalmente, un *Proceso de Decisión de Markov Finito* (MDP) se define mediante la tupla:

$$\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle, \tag{2.15}$$

donde:

- $\mathcal{S}$ es el conjunto finito de *estados* posibles del entorno.
- $\mathcal{A}$ es el conjunto finito de *acciones* disponibles para el agente.
- P es la *función de probabilidad de transición*, definida como

  $$P(s'|s,a) = \Pr\{S_{t+1} = s' \,|\, S_t = s,\, A_t = a\}, \tag{2.16}$$

  la cual representa la probabilidad de que el entorno transite desde el estado s al estado s' tras ejecutar la acción a.
- R es la *función de recompensa esperada*, dada por

  $$R(s,a) = \mathbb{E}[R_{t+1} \,|\, S_t = s,\, A_t = a], \tag{2.17}$$

  que indica el valor medio de la recompensa recibida al aplicar la acción a en el estado s.
- $\gamma \in [0,1]$ es el *factor de descuento*, que pondera la importancia relativa de las recompensas futuras con respecto a las inmediatas.

Otro concepto importante dentro del Aprendizaje por Refuerzo es el de *episodio* (*episode*), entendido como una secuencia completa de interacciones agente-entorno que comienza en un estado inicial y finaliza al alcanzar un *estado terminal*, definido por una condición de término. Dichas condiciones pueden variar según los objetivos del entrenamiento. Una vez finalizado un episodio, el entorno se reinicia ya sea a un estado inicial fijo o a una muestra proveniente de una distribución de estados iniciales.

Por otro lado, existen situaciones en las cuales la interacción agente-entorno no presenta un final natural ni episodios claramente identificables, permitiendo que la tarea continúe indefinidamente. Este tipo de formulación se usa problemas de control de procesos en operación continua [12].

El objetivo del agente en un Proceso de Decisión de Markov es encontrar una estrategia de comportamiento, o *política*, que especifique la acción a ejecutar en cada estado y que maximice

la recompensa acumulada esperada a lo largo del tiempo. La política que logra este máximo se denomina *política óptima* y suele representarse como π^*.

En términos generales, una política puede expresarse de manera estocástica como una distribución de probabilidad sobre las acciones, es decir, $\pi(a \mid s) = \Pr\{A_t = a \mid S_t = s\}$. Bajo una política dada π, el *valor de estado* $V^\pi(s)$ representa la recompensa acumulada esperada que el agente puede obtener al partir del estado s y seguir dicha política [12]. Esta relación puede expresarse mediante la *ecuación de Bellman*:

$$V^\pi(s) = \sum_a \pi(a|s) \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^\pi(s') \right], \tag{2.18}$$

donde $R(s,a)$ es la recompensa inmediata esperada, $\gamma \in [0,1]$ es el factor de descuento y $P(s'|s,a)$ la probabilidad de transición del entorno. En términos intuitivos, el agente espera recibir la recompensa inmediata $R(s,a)$ por realizar la acción recomendada por la política π, y transitar hacia un estado futuro con un valor esperado $V^\pi(s')$ ponderado por γ.

De acuerdo con la teoría de la programación dinámica [13], existe al menos una política óptima estacionaria π^* que maximiza el valor esperado de cada estado. El *valor óptimo* $V^*(s)$ se define entonces como:

$$V^*(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s'|s,a)\, V^*(s') \right]. \tag{2.19}$$

El objetivo del aprendizaje por refuerzo es encontrar una *política óptima* π^* que maximice la recompensa total esperada a lo largo del tiempo. Una política π define una estrategia de decisión que el agente sigue, especificando la probabilidad de seleccionar cada acción a dado un estado s, es decir, $\pi(a|s) = \Pr\{A_t = a \,|\, S_t = s\}$. Entre todas las políticas posibles, la política óptima π^* es aquella que genera el mayor valor esperado en todos los estados del entorno:

$$V^*(s) = \max_\pi V^\pi(s), \qquad \forall s \in \mathcal{S}. \tag{2.20}$$

El *principio de optimalidad* formulado por Bellman establece que, independientemente del estado inicial y de la decisión tomada, las decisiones subsecuentes deben conformar una política óptima respecto al estado que resulte de la primera acción. Esto implica que, si una secuencia de decisiones es óptima desde el inicio, entonces la subsecuencia resultante también lo será desde cualquier estado intermedio alcanzado bajo esa política [13]. Formalmente, este principio puede expresarse mediante la *ecuación de optimalidad de Bellman*:

$$V^*(s) = \max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V^*(s') \right], \tag{2.21}$$

la cual expresa que el valor óptimo de un estado corresponde a la recompensa inmediata más el valor esperado del mejor estado futuro, ponderado por el factor de descuento γ.

2.5. Algoritmos basados en aprendizaje por refuerzo

El *reinforcement learning* incluye distintas familias de métodos para resolver problemas de control, como la programación dinámica, los métodos Monte Carlo y las técnicas basadas en *temporal-difference* (TD). A partir de los fundamentos teóricos previamente presentados, donde se estableció

el marco de los procesos de decisión de Markov y las ecuaciones de Bellman, nacen algoritmos que permiten a un agente aprender directamente de la interacción con su entorno. Estos algoritmos buscan estimar las funciones de valor $V^*(s)$ sin requerir conocer explícitamente el modelo de transición $P(s'|s,a)$.

Dentro de los métodos basados en TD existen enfoques *on-policy*, como Sarsa, que evalúan y mejoran la misma política que ejecutan. En cambio, el método en el que se centra este documento, Q-learning, aprende de manera independiente de la política utilizada durante la exploración. El nombre del algoritmo proviene de la función $Q(s,a)$, donde la letra Q se asocia con *quality*, refiriéndose a la calidad de ejecutar una acción determinada en un estado dado Watkins y Dayan. Esto significa que el agente puede comportarse de forma exploratoria mientras aprende, simultáneamente, la política óptima. La razón es que Q-learning no estima el valor esperado de seguir la política actual, sino que actualiza directamente una aproximación de la función de acción-valor óptima, denotada como Q^*.

Una de las fortalezas más importantes de Q-learning es su garantía de convergencia bajo ciertas condiciones. Si todas las parejas estado-acción se visitan con suficiente frecuencia y la tasa de aprendizaje se ajusta adecuadamente, las estimaciones de Q convergen con probabilidad uno a la función óptima Q*. Este resultado permitió demostrar que es posible obtener control óptimo aun sin conocer la dinámica del sistema y mientras se explora el entorno [14].

Gracias a estas características, Q-learning se transformó en uno de los algoritmos más influyentes y utilizados para resolver problemas de control en entornos discretos, especialmente cuando el sistema no cuenta con un modelo claro o su dinámica es difícil de describir[12].

El algoritmo Q-learning, al ser tabular, requiere un espacio de estados discretizado para poder aprender del entorno. Esto es especialmente relevante desde la perspectiva del esfuerzo computacional, ya que el tamaño de la matriz Q depende directamente del nivel de discretización necesario para el control y del rango total de los estados. Esta limitación provoca que, cuando el espacio de estados es muy grande, el algoritmo se vuelva costoso a nivel computacional debido a la gran cantidad de valores Q que deben actualizarse en cada iteración.

Para superar esta restricción, es posible recurrir a técnicas como el *Deep Q-Network* (DQN), una extensión que utiliza una red neuronal artificial capaz de aproximar la función $Q(S,A)$. En una DQN, la tabla es reemplazada por una red neuronal que recibe el estado s y produce como salida un vector con los valores $Q(s,a)$ para cada acción a. Esto permite manejar espacios de estado de mayor dimensión e incluso continuos, a diferencia del Q-learning tabular, que solo opera sobre espacios discretos [15].

2.5.1. Q-Learning

Como se mencionó anteriormente, el Q-learning es un método de Aprendizaje por Refuerzo basado en diferencias temporales (TD), cuyo objetivo es aprender la acción óptima a ejecutar en cada estado sin requerir conocimiento previo del modelo dinámico del sistema. La idea central del algoritmo consiste en actualizar, de manera iterativa, una tabla de valores denominada $Q(s,a)$, la cual representa el valor estimado de ejecutar la acción a en el estado s. A esta función se le conoce como *función de acción-valor*. El algoritmo fue planteado por Watkins y Dayan en «Technical Note Q-Learning. Machine Learning»[14]:

$$Q(S_t,A_t) \leftarrow Q(S_t,A_t) + \alpha\big[R_{t+1} + \gamma \max_a Q(S_{t+1},a) - Q(S_t,A_t)\big], \tag{2.22}$$

Esta ecuación corresponde a la regla de actualización del *Q-learning*, donde cada término cumple un rol fundamental en el proceso de aprendizaje. El valor $Q(S_t, A_t)$ representa la estimación actual del valor de ejecutar la acción A_t en el estado S_t, mientras que $Q(S_{t+1}, a)$ denota los valores estimados para las posibles acciones a en el siguiente estado S_{t+1}.

El parámetro α es la *tasa de aprendizaje*, que determina cuánto influye la nueva información sobre el valor previamente almacenado. El término R_{t+1} corresponde a la *recompensa inmediata* obtenida tras realizar la acción. Por su parte, el parámetro $\gamma \in [0, 1)$ es el *factor de descuento*, que establece la importancia relativa de las recompensas futuras (a mayor valor de γ, mayor relevancia tienen dichas recompensas).

La expresión $\max_a Q(S_{t+1}, a)$ identifica el mejor valor estimado para el próximo estado entre las acciones disponibles. Finalmente, el término completo dentro de los corchetes constituye el *error temporal*, que cuantifica la diferencia entre la estimación actual y la nueva estimación basada en la experiencia. Dado la que la formulación del Q-learning viene dada como un proceso de decisión de Markov, se puede esquematizar la interacción agente-entorno, tal como muestra la Figura 2.4.

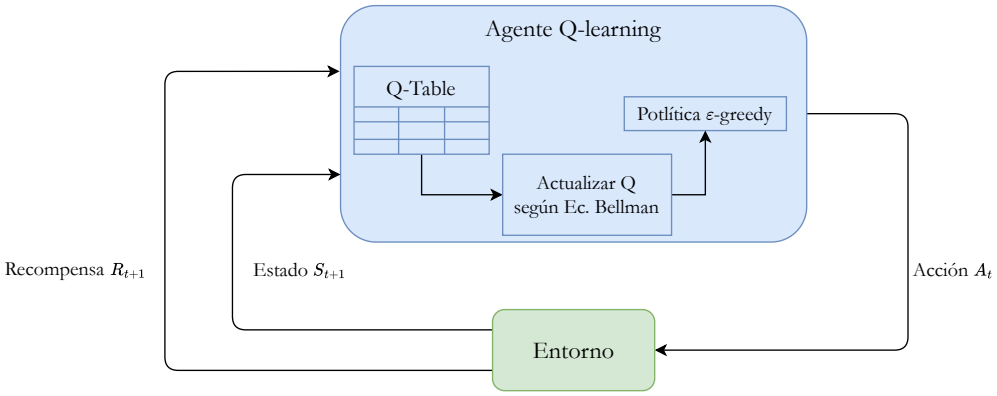


Figura 2.4: Esquema de Q-learning como MDP.

Cada vez que el agente ejecuta una acción y observa la recompensa y el siguiente estado, ajusta el valor almacenado en la tabla mediante la regla de actualización mostrada en la Ecuación 2.22. Este proceso iterativo guía al agente hacia la estimación de la función de acción-valor óptima, denotada como Q^*, la cual permite identificar la mejor acción en cada estado.

Para garantizar la exploración, Q-learning utiliza típicamente una política ε-greedy, que selecciona acciones aleatorias con baja probabilidad mientras sigue mayoritariamente la acción con el mayor valor estimado. Debido a esto, Q-learning se clasifica como un método *fuera de política* (*off-policy*), ya que aprende la política óptima aun cuando las acciones ejecutadas durante la exploración no corresponden necesariamente a aquellas que maximizan la recompensa inmediata [12, 14].

Considerando un valor $\varepsilon < 1$, la política ε-greedy se define como:

$$\varepsilon\text{-greedy} = \begin{cases} 1-\varepsilon & \text{seleccionar la acción } A_t = \arg\max_a Q(S_t, a) \\ \varepsilon & \text{seleccionar una acción } A_t \text{ aleatoria.} \end{cases} \tag{2.23}$$

Debido a que la función Q se almacena en forma tabular, Q-learning requiere discretizar los estados del entorno. Esto lo convierte en un método sencillo y eficiente cuando el espacio de estados es reducido o puede discretizarse de manera adecuada, como ocurre en ciertos problemas de control en electrónica de potencia.

Otra cualidad importante del algoritmo de aprendizaje Q-learning es que, al actualizar los valores Q en cada iteración, permite realizar un control *online* del entorno. En otras palabras, el agente puede ir aprendiendo y ajustando su política a medida que interactúa con el sistema, actualizando sus estimaciones en cada paso de muestreo sin requerir previamente un proceso de entrenamiento.

Este enfoque resulta especialmente útil en tareas de control que no presentan episodios con un término natural, como ocurre en sistemas en operación continua. En estos casos, la interacción agente-entorno puede prolongarse indefinidamente y el algoritmo sigue pudiendo aprender directamente de la experiencia mientras controla el sistema, adaptándose de inmediato a las dinámicas que observa.

El algoritmo Q-learning se ejecuta de manera iterativa siguiendo los pasos que se describen a continuación. Esta versión corresponde a una adaptación del pseudocódigo original propuesto por Sutton y Barto [12], ajustado para representar el proceso general de aprendizaje en un entorno de control.

Algoritmo 2.1 Q-learning para control en un entorno genérico [12]

Inicializar arbitrariamente la tabla $Q(s, a)$ para todos los estados s y acciones a
para cada episodio **hacer**
    Inicializar el estado inicial S_0
    para cada paso de tiempo t **hacer**
        Seleccionar A_t a partir de S_t usando una política de comportamiento (ε-greedy 2.23)
        Aplicar la acción A_t al entorno
        Observar la recompensa R_{t+1} y el nuevo estado S_{t+1}
        Actualizar el valor:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\big]$$

        Definir $S_t \leftarrow S_{t+1}$
        si S_t es un estado terminal **entonces**
            Terminar Episodio

2.5.2. Deep Q-Network (DQN)

La formulación tabular de Q-learning permite aproximar la función de acción-valor $Q(s, a)$ mediante una tabla que se actualiza con la Ecuación 2.22. Sin embargo, este enfoque se vuelve impracticable cuando el espacio de estados es continuo, de alta dimensión o requiere representar relaciones complejas entre variables, como ocurre en sistemas dinámicos, tareas de control no lineales o entornos con observaciones sensoriales ruidosas. Una alternativa ampliamente utilizada es aproximar esta función mediante un modelo paramétrico. Dentro de estos modelos, las redes neuronales destacan por su capacidad para aproximar funciones no lineales complejas [16].

Una red neuronal consiste en un conjunto de capas de unidades conectadas entre sí, mediante pesos ajustables. Dados unos parámetros θ (que incluyen pesos y sesgos de todas las capas), la red se define como una función $f(x, \theta)$ que transforma una entrada x en una salida. Durante el entrenamiento, lo parámetros θ se modifican mediante un algoritmo de optimización para que la salida de la red se acerque a un valor deseado [17].

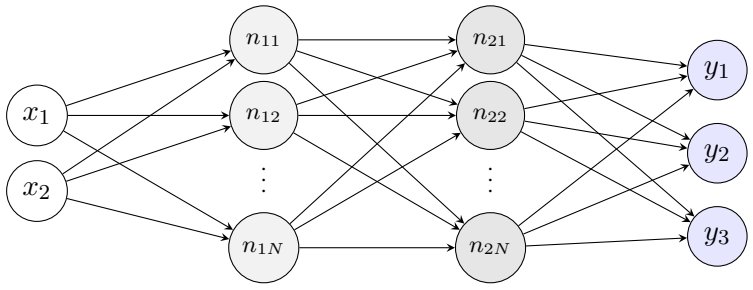


Figura 2.5: Esquema genérico de una red neuronal artificial con N neuronas en cada capa oculta.

La Figura 2.5 muestra un esquema conceptual de una red neuronal, construido para ilustrar la organización típica de sus capas. La capa de entrada corresponde al conjunto de observaciones utilizadas para entrenar la red. La cantidad de capas y neuronas ocultas define la arquitectura del modelo, la cual puede ajustarse según los requerimientos del problema. Cada neurona posee un conjunto de entradas asociadas a pesos y un sesgo, parámetros que determinan las operaciones necesarias para generar una salida a partir de las observaciones. La capa de salida representa la respuesta de la red neuronal frente al estado actual de las neuronas en las capas ocultas. En este tipo de redes, la información fluye secuencialmente desde la capa de entrada hacia la de salida, sin retroalimentación entre neuronas.

Una neurona artificial es la unidad fundamental de una red neuronal y opera realizando una combinación lineal de sus entradas seguida de una transformación no lineal. Cada neurona recibe un conjunto de entradas x_i, que son ponderadas mediante coeficientes w_i y sumadas junto con un término de sesgo b. El resultado de esta combinación lineal, $\sum_i w_i x_i + b$, es posteriormente procesado por una función de activación $f(\cdot)$, la cual introduce no linealidad y permite que la red neuronal aprenda relaciones complejas entre las variables de entrada. De esta forma, la salida de una neurona se expresa como $y = f\left(\sum_i w_i x_i + b\right)$, representando una transformación elemental pero esencial para la construcción de modelos de aprendizaje profundo [17].

Aprovechando la capacidad de las redes neuronales como aproximadores universales, Mnih et al. propusieron en «Playing Atari with Deep Reinforcement Learning» [18] una extensión de Q-learning que reemplaza la tabla por una red neuronal profunda capaz de aproximar $Q(s, a)$ para cualquier par estado-acción. Este método lleva por nombre *Deep Q-Network* (DQN), y su objetivo es lograr que el agente aprenda una función $Q(s, a; \theta)$ parametrizada, que se aproxime el valor óptimo Q^*. Esta nueva función aproximadora se denomina *Q-Network*. Dado que el agente es reemplazado por una red neuronal, se puede esquematizar el MDP como se muestra en la Figura 2.6.

En algoritmos de aprendizaje por refuerzo al utilizar aproximaciones no lineales de la función $Q(s, a)$, como una red neuronal, el aprendizaje se puede volver inestable o incluso divergir, esto puede deberse por correlación temporal presentadas en la secuencia de observación, pequeñas actualizaciones de Q puede cambiar significativamente la política y por lo tanto, la distribución de los datos [18, 19].

Para abordar estos problemas, Mnih et al. introducen dos técnicas clave. En primer lugar, implementan un mecanismo llamado *repetición de experiencias* (*experience replay*), que aleatoriza las muestras de datos, eliminando la correlación temporal entre observaciones y suavizando

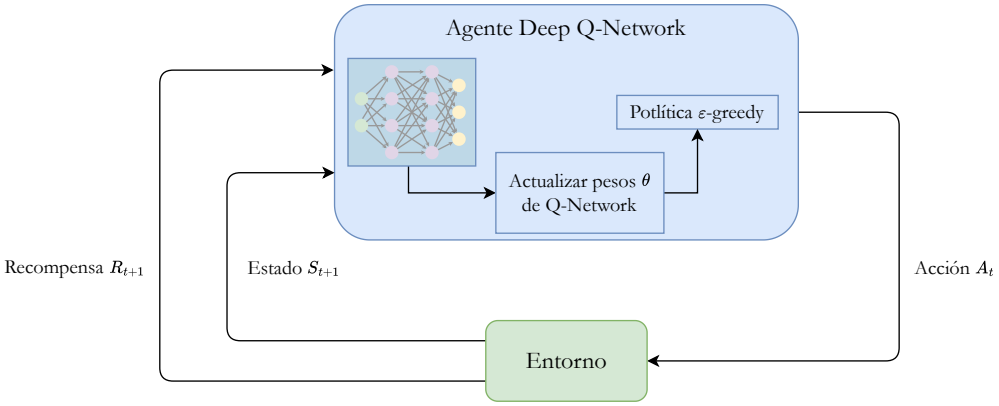


Figura 2.6: Esquema de Deep Q-Network como MDP.

los cambios en la distribución de datos. En segundo lugar, se emplea una red Q objetivo, cuya actualización periódica ajusta la función de valor Q hacia los valores óptimos de manera más estable.

Formalmente, la técnica de *experience replay* almacena en cada paso de tiempo t experiencias de la forma $e_t = (S_t, A_t, R_t, S_{t+1})$ en un conjunto de datos $\mathcal{D} := \{e_1, \ldots, e_t\}$ recopilado a partir de múltiples episodios y almacenado en una *memoria de repetición* (*replay buffer*), que guarda las experiencias de los últimos t pasos de tiempo. Durante el aprendizaje, la función Q se actualiza a partir de pequeñas muestras aleatorias de experiencias, $\mathcal{M} : (S, A, R, S_{t+1}) \sim U(\mathcal{D})$ (*minibatches*), extraídas uniformemente al azar desde la memoria en lugar de usar transiciones consecutivas. Es importante destacar que esta técnica es compatible con el algoritmo Q-learning debido a su naturaleza *off-policy*, ya que los parámetros actuales pueden diferir de aquellos usados para generar las muestras.

La segunda modificación mencionada busca mejorar aún más la estabilidad del aprendizaje al utilizar redes neuronales. Para ello, se emplean dos redes separadas: la aproximación de la función de acción-valor, Q-Network $Q(s, a; \theta)$, que se actualiza en cada paso de tiempo; y la red objetivo, $\hat{Q}$-Network $Q(s, a; \theta^-)$, cuyos parámetros θ^- se mantienen congelados durante C pasos antes de actualizarse mediante $\theta^- \leftarrow \theta$. Los valores objetivo y, definidos a partir de la red objetivo $\hat{Q}$, se actualizan mediante la ecuación de Bellman:

$$y = r + \gamma \max_{a'} \hat{Q}(S_{t+1}, a'; \theta^-). \tag{2.24}$$

La forma más habitual de medir el rendimiento del modelo consiste en calcular el *error cuadrático medio* (MSE) en un conjunto de prueba. En *Deep Learning*, dicho criterio se formaliza mediante una función de pérdida. En este caso, el comportamiento del modelo está dado por la Q-Network, mientras que los valores objetivo provienen de la $\hat{Q}$-Network. Durante el entrenamiento, se aplican actualizaciones a $Q(s, a; \theta)$ utilizando minibatches de experiencia $\mathcal{M}$ muestreados uniformemente al azar desde la memoria de experiencias $\mathcal{D}$ [17, 20]. Así, la función de pérdida en cada iteración i se define como:

$$L_i = \mathbb{E}_{\mathcal{M}} \left[(y_i - Q(s, a; \theta_i))^2 \right]. \tag{2.25}$$

Para minimizar el valor de la función de pérdida L_i, y por ende el *MSE*, es necesario encontrar

valores de θ que lo reduzcan. A partir de la ecuación 2.25, el gradiente correspondiente se expresa como:

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}_{\mathcal{M}} \left[(y_i - Q(s, a; \theta_i)) \nabla_{\theta_i} Q(s, a; \theta_i) \right]. \tag{2.26}$$

Una vez definida la función de pérdida y su correspondiente gradiente respecto de los parámetros de la red, es posible describir el mecanismo mediante el cual el modelo ajusta sus pesos durante el entrenamiento. En particular, el procedimiento estándar consiste en desplazarse en la dirección opuesta al gradiente, con el fin de minimizar el valor de la función de pérdida [21]. Este proceso se formaliza mediante la siguiente regla de actualización:

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} L(\theta), \tag{2.27}$$

donde α representa la tasa de aprendizaje que controla la magnitud del paso dado en cada iteración.

En la práctica, la esperanza matemática mostrada en la ecuación 2.25 se aproxima mediante un promedio sobre un minibatch $\mathcal{M}$ de tamaño $M = |\mathcal{M}|$. De este modo, la función de pérdida puede escribirse de manera equivalente como:

$$L(\theta) = \frac{1}{2M} \sum_{(S_j, A_j, R_j, S_{j+1}) \in \mathcal{M}} \big(y_j - Q(S_j, A_j; \theta)\big)^2, \tag{2.28}$$

En el Algoritmo 2.2, se presenta un esquema genérico de DQN, adaptado de lo propuesto por Mnih et al. Originalmente, el algoritmo está orientado al aprendizaje de agentes en entornos de videojuegos, donde las observaciones corresponden a secuencias de imágenes procesadas mediante redes convolucionales.

Algoritmo 2.2 Deep Q-Network con *experience replay*

Inicializar la memoria de repetición $\mathcal{D}$ con capacidad N
Inicializar la función de acción-valor $Q(s, a; \theta)$ con pesos aleatorios θ
Inicializar la red objetivo $\hat{Q}(s, a; \theta^-)$ con $\theta^- \leftarrow \theta$
para cada episodio **hacer**
    Inicializar el estado inicial S_1 del entorno
    para cada paso de tiempo t **hacer**
        Con probabilidad ε seleccionar una acción aleatoria $A_t \in \mathcal{A}$
        En caso contrario, seleccionar

$$A_t = \arg\max_a Q(S_t, a; \theta)$$

        Ejecutar A_t en el entorno y observar la recompensa R_t y el siguiente estado S_{t+1}
        Almacenar la transición (S_t, A_t, R_t, S_{t+1}) en la memoria $\mathcal{D}$
        Muestrear aleatoriamente un *minibatch* de transiciones

$$\mathcal{M} = \{(S_j, A_j, R_j, S_{j+1})\} \sim U(\mathcal{D})$$

        para cada transición $(S_j, A_j, R_j, S_{j+1}) \in \mathcal{M}$ **hacer**
            Calcular el valor objetivo

$$y_j = \begin{cases} R_j, & \text{si episodio termina en } j+1 \\ R_j + \gamma \max_a \hat{Q}(S_{j+1}, a; \theta^-), & \text{en caso contrario} \end{cases}$$

        Actualizar los parámetros de la Q-Network: $\theta \leftarrow \theta - \alpha \nabla_\theta L(\theta)$
        Cada C pasos de tiempo, actualizar los parámetros de la red objetivo $\hat{Q}$:

$$\theta^- \leftarrow \theta$$

En DQN, la estrategia de exploración utilizada corresponde a una política ε-greedy con *decaimiento*. En este esquema, el agente selecciona acciones aleatorias con probabilidad ε y, con probabilidad $1 - \varepsilon$, ejecuta la acción que maximiza el valor estimado por la red. A diferencia del caso tabular, donde se empleó una exploración fija, en DQN el valor de ε disminuye progresivamente durante el entrenamiento. Esto permite comenzar con una exploración amplia, necesaria para recopilar experiencias diversas, y avanzar gradualmente hacia un comportamiento más explotador a medida que la red mejora sus estimaciones. Herramientas como *Reinforcement Learning Toolbox* de *MATLAB* [22] calculan el valor de ε que decae en cada paso de tiempo según

$$\varepsilon_{\text{k+1}} = \varepsilon_{\text{k}} \cdot (1 - \varepsilon_{\text{decay}}) \tag{2.29}$$

Otra técnica utilizada por este algoritmo es *Double DQN (DDQN)*, la cual contribuye significativamente a estabilizar el proceso de aprendizaje en redes de valor. En lugar de emplear la misma red para decidir la mejor acción y estimar su valor, DDQN utiliza la red principal para seleccionar la acción y la red objetivo para evaluarla, mitigando así la sobreestimación sistemática de los valores Q. Esta separación disminuye la volatilidad en las actualizaciones, produce objetivos de aprendizaje más consistentes y mejora la estabilidad general del entrenamiento, lo que se traduce en una convergencia más confiable y un desempeño superior en comparación con el DQN

convencional [23]. Este procedimiento culmina con el reemplazo del valor objetivo mostrado en la Ecuación 2.24, por la siguiente expresión:

$$y = r + \gamma \max_{a} \hat{Q}(S_{t+1}, \arg\max_{a} Q(S_{t+1}, a; \theta); \theta^{-}). \tag{2.30}$$

Por otro lado, en el enfoque original propuesto por Mnih et al., la actualización de la red objetivo corresponde a un método de *actualización periódica* (*periodic update*). Tal como se describió anteriormente, este esquema consiste en copiar los parámetros de la red principal (Q-Network) hacia la red objetivo ($\hat{Q}$-Network) cada C pasos de tiempo, manteniendo valores constantes entre dichas actualizaciones para mejorar la estabilidad del entrenamiento.

Una alternativa ampliamente utilizada en la literatura es la *actualización suave* (*smoothing* o *soft target update*), donde los parámetros de la red objetivo se ajustan gradualmente en cada iteración a través de un promedio ponderado de los parámetros actuales y anteriores [24]. Esta actualización se expresa mediante

$$\theta^{-} \leftarrow \tau\, \theta + (1 - \tau)\, \theta^{-}, \tag{2.31}$$

donde $0 < \tau \ll 1$, denominado *Target Smooth factor*, controla la rapidez con que la red objetivo sigue a la red en principal. Este mecanismo reduce las oscilaciones asociadas a las actualizaciones abruptas y constituye la base de los métodos de suavizado implementados en diversas bibliotecas de aprendizaje por refuerzo.

El entrenamiento de un agente DQN resulta complejo debido a la alta sensibilidad del algoritmo a la elección de sus hiperparámetros y a la naturaleza inestable del aprendizaje con funciones aproximadoras profundas. Factores como la tasa de aprendizaje, el tamaño del *replay buffer*, el *target update*, el diseño de la red y el esquema de exploración pueden influir significativamente en la capacidad del agente para converger hacia la política deseada. Además, la correlación temporal de las muestras y la propagación de errores en la estimación de los valores Q pueden provocar inestabilidad, oscilaciones o incluso divergencia si la configuración no es adecuada. Estas dificultades han sido ampliamente documentadas en la literatura de *deep reinforcement learning*, donde se destaca que encontrar una combinación apropiada de parámetros es clave para lograr un entrenamiento estable y una política efectiva [21].

Capítulo 3

Configuración del Sistema y de los Agentes de Aprendizaje

En los capítulos anteriores se presentó la formulación general del problema de control como un Proceso de Decisión de Markov y se definieron los entornos de aprendizaje utilizados por los algoritmos Q-learning y Deep Q-Network (DQN). Sobre esta base, el presente capítulo describe los casos de estudio específicos considerados en este trabajo, así como la configuración experimental empleada para cada uno de ellos.

El objetivo de este capítulo es establecer con claridad las condiciones bajo las cuales se entrenarán y evaluarán los agentes de aprendizaje por refuerzo, detallando las variantes monofásicas y trifásicas del problema, las arquitecturas adoptadas, las discretizaciones utilizadas y las plataformas de simulación asociadas. Estas descripciones permiten contextualizar los experimentos del capítulo siguiente y evitar ambigüedades al comparar el desempeño de los diferentes métodos.

En particular, se presentan los casos monofásicos y trifásicos tanto para Q-learning como para DQN, indicando los parámetros relevantes para cada escenario, las características propias de cada entorno de simulación y la estructura general de los agentes.

3.1. Modelo eléctrico del sistema

El diseño y la evaluación de los agentes de aprendizaje por refuerzo se realizan sobre un sistema eléctrico común para todos los casos estudiados. Este capítulo describe los parámetros físicos utilizados en las simulaciones, los cuales permanecen invariantes tanto para el agente Q-learning como para el agente DQN, de modo que cualquier diferencia observada en el desempeño se deba exclusivamente al método de control.

3.1.1. Sistema monofásico

El caso monofásico corresponde a un inversor monofásico de dos niveles con topología en puente completo (Figura 2.1). El convertidor está compuesto por cuatro interruptores IGBT con sus respectivos diodos en antiparalelo, dispuestos en dos ramas que permiten aplicar a la carga los niveles de tensión $+V_{\mathrm{dc}}$, 0 y $-V_{\mathrm{dc}}$ según el par de interruptores activados.

Los parámetros eléctricos base del sistema utilizados en las simulaciones son los siguientes:

$$V_{\mathrm{dc}} = 200\ \mathrm{V}, \qquad R = 5\ \Omega, \qquad L = 10\ \mathrm{mH}. \tag{3.1}$$

La tensión aplicada a la carga se obtiene a partir del estado de conmutación del puente completo, lo que permite generar la corriente necesaria para seguir la referencia impuesta por el controlador. Esta topología es representativa de convertidores monofásicos ampliamente utilizados en aplicaciones de control de corriente, y constituye un escenario adecuado para evaluar el comportamiento del agente Q-learning y del agente DQN en una estructura física realista, manteniendo una complejidad moderada.

3.1.2. Sistemas trifásicos

Los casos trifásicos comparten la misma estructura de inversor de dos niveles, extendida a tres ramas. En la Figura 3.1, se muestra el esquema eléctrico correspondiente al inversor trifásico en donde $S_p,\ p \in \{a, b, c\}$ corresponden las señales de disparo de las fases correspondientes, y $\overline{S_p}$ a la señal de la fase respectiva.

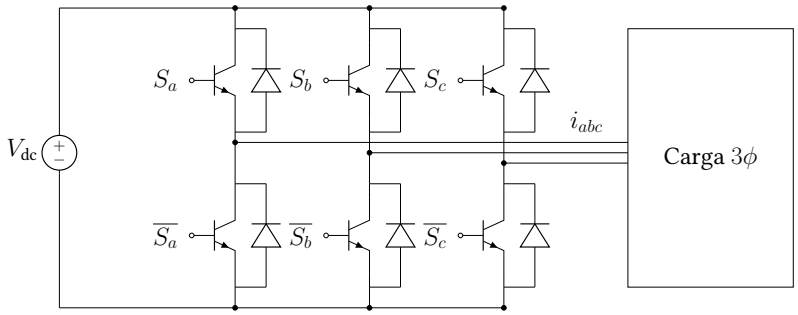


Figura 3.1: Esquema eléctrico de inversor de dos niveles trifásico conectado a carga trifásica.

Dependiendo del escenario considerado, la carga puede corresponder a un conjunto RL trifásico o a un motor de inducción trifásico modelado en su representación estándar. Los parámetros eléctricos específicos de estos casos se detallan en las secciones correspondientes.

3.2. Formulación del Proceso de Decisión de Markov

El diseño del controlador basado en aprendizaje por refuerzo requiere formular el problema de control como un Proceso de Decisión de Markov (MDP). Esta representación define los elementos fundamentales de la interacción entre el agente y su entorno: el conjunto de estados, las acciones disponibles, la función de transición y la función de recompensa. Aunque el sistema eléctrico que se busca controlar es el mismo, la forma en que estos elementos se representan difiere entre Q-learning y *Deep Q-Network* (DQN), debido a las particularidades de cada enfoque.

3.2.1. Conjunto de estados

Q-learning: estado discreto

En el caso del Q-learning tabular, la elección de las variables que compondrán el estado es un aspecto crítico, ya que determina directamente el tamaño de la tabla $Q(s, a)$ y, por tanto, la factibilidad del método. Dado que la tabla debe almacenar un valor para cada par (s, a), la cardinalidad del conjunto de estados tiene un impacto lineal en los tiempos de aprendizaje, en el almacenamiento requerido y en la estabilidad del proceso.

Un diseño ingenuo podría considerar como estado tanto la corriente medida como la corriente de referencia, discretizadas de manera independiente. Si se emplearan N_{ref} niveles para la referencia y N_{meas} niveles para la corriente medida, el tamaño total del espacio de estados sería

$$|S| = N_{\text{ref}} \times N_{\text{meas}}. \tag{3.2}$$

Si se discretiza cada corriente, por ejemplo $N_{\text{ref}} = N_{\text{meas}} = 200$, según lo considerado en la Ecuación 3.2, el número de estados alcanzaría $|S| = 40\,000$, lo que implica que la tabla $Q(s, a)$ tendría $40\,000 \times |\mathcal{A}|$ entradas. Con tres acciones posibles, esto corresponde a más de $120\,000$ valores que deben actualizarse durante cada paso de simulación. Este tamaño resulta prohibitivo para un controlador que opera a una frecuencia de muestreo del orden de decenas de microsegundos, generando tiempos de cómputo excesivos y dificultando la convergencia del aprendizaje.

Por esta razón, en este trabajo se emplea una representación discreta del estado basada en el error instantáneo de corriente:

$$s[k] = f_{\text{disc}}(i_{\text{ref}}(t) - i(t)) = f_{\text{disc}}(e(t)), \tag{3.3}$$

en donde $f_{\text{disc}}(\cdot)$, asigna el error de corriente a un estado discretizado con una resolución e_{step}. Esta decisión reduce de manera drástica el tamaño del espacio de estados,

$$|S| = N_e, \tag{3.4}$$

donde N_e es del orden de cientos, dependiendo del grado de discretización. De esta manera, la tabla $Q(s, a)$ resulta factible de actualizar en cada paso de simulación, permitiendo implementar el aprendizaje de forma completamente *online* sin comprometer el desempeño temporal del controlador.

Esta reducción dimensional no solo mejora la eficiencia computacional, sino que además evita la redundancia inherente a representar por separado variables cuya relación funcional es conocida, favoreciendo una estructura de estado más compacta y adecuada para el método tabular.

DQN: estado continuo

A diferencia del enfoque tabular, DQN trabaja directamente con estados continuos. Para el sistema monofásico, el vector de observación se define como

$$s[k] = \begin{bmatrix} i[k] \\ i_{\text{ref}}[k] \end{bmatrix}. \tag{3.5}$$

En el caso trifásico, el estado incluye las tres corrientes medidas además de las corrientes de referencia, según el sistema abc, $\alpha\beta$ o dq.

Esta representación continua permite capturar relaciones no lineales y utilizar directamente las capacidades de aproximación de las redes neuronales profundas.

3.2.2. Conjunto de acciones

El conjunto de acciones corresponde a los posibles niveles de tensión que el inversor puede aplicar. Para el caso monofásico se tiene que las tensiones aplicables a la carga son V=V_{dc},0,$-V_{\text{dc}}$, las cuales se generan dependiendo de las señales de compuerta de los IGBT S_1 y S_2, como se muestra en la Figura 2.1. Por lo que las acciones posibles corresponden a:

$$a \in \{(1,0),(0,0),(0,1)\}. \tag{3.6}$$

Se omite la acción $a = (1,1)$ ya que produce el mismo nivel de tensión que $a = (0,0)$ y es necesario mantener el nivel dimensional de Q al mínimo.

En el caso trifásico, las acciones del agente se asocian directamente a los vectores de conmutación válidos del inversor de dos niveles. Dado que cada una de las tres fases puede adoptar de forma independiente un estado binario, existen en total $2^3 = 8$ combinaciones posibles del vector de conmutación (S_a, S_b, S_c). Las combinaciones $(0,0,0)$ y $(1,1,1)$ generan el mismo vector espacial nulo, pero corresponden a configuraciones de conmutación físicamente distintas, ya que todas las fases están conectadas al polo negativo o al polo positivo de la fuente, por lo que ambas deben considerarse en el conjunto de acciones debido a sus implicancias dinámicas.

Por lo tanto, el conjunto de acciones para el inversor de dos niveles trifásico viene determinado por lo mostrado en la Tabla 3.1.

Tabla 3.1: Vectores de conmutación de un inversor trifásico de dos niveles

Vector	V_0	V_1	V_2	V_3	V_4	V_5	V_6	V_7
$(S_a\, S_b\, S_c)$	(000)	(100)	(110)	(010)	(011)	(001)	(101)	(111)

Tanto en el caso monofásico como trifásico, el conjunto acciones es común a Q-learning y DQN, ya que ambos agentes aprenden políticas basadas en las mismas posibilidades de conmutación.

3.2.3. Función de recompensa

La función de recompensa constituye el mecanismo mediante el cual el agente evalúa la calidad de sus acciones y orienta su proceso de aprendizaje hacia el objetivo del control. En el presente trabajo, dicho objetivo consiste en lograr un seguimiento adecuado de la corriente de referencia, manteniendo una dinámica estable.

Independientemente del algoritmo empleado, la recompensa se diseña para penalizar el error de seguimiento, de modo que acciones que acerquen la corriente medida a su referencia reciban un valor mayor (menos negativo), mientras que desviaciones significativas sean penalizadas. Esta estructura general permite que tanto los métodos tabulares como aquellos basados en aproximadores neuronales ajusten su política en función de la calidad del seguimiento de corriente.

Adicionalmente, en los casos en que se requiera, la recompensa puede incorporar términos asociados a otros objetivos relevantes, tales como limitar variaciones abruptas de la acción o moderar el número de conmutaciones, manteniendo siempre como criterio central la minimización del error de corriente. Aunque la formulación explícita difiere entre Q-learning y DQN, ambos enfoques comparten esta lógica común orientada al desempeño del sistema.

3.2.4. Episodios de aprendizaje

Tal como se explicó en la Sección 2.4, un *episodio* en Aprendizaje por Refuerzo corresponde a una secuencia completa de interacciones agente–entorno, delimitada por un estado inicial y una condición de término explícita. No obstante, algunos problemas de control no presentan episodios naturalmente definidos, permitiendo que la interacción continúe de manera indefinida. Esta distinción resulta particularmente relevante al comparar el comportamiento de los algoritmos implementados en este trabajo.

En el caso de Q-learning, el agente opera de manera *online* y en un entorno continuo, sin un estado terminal intrínsecamente definido. Por esta razón, el episodio coincide con la duración total de la simulación utilizada para evaluar el desempeño del controlador. En cada paso de tiempo, el agente observa el estado, selecciona una acción y actualiza la función $Q(s, a)$, de modo que el aprendizaje ocurre de manera continua, ejecutando el procedimiento descrito en el Algoritmo 2.1, sin segmentación explícita en episodios ni necesidad de reiniciar el entorno.

Por el contrario, el enfoque DQN requiere trabajar con episodios claramente definidos para permitir el uso de memoria de repetición y entrenamiento por lotes. En este caso, cada episodio corresponde al seguimiento de una referencia de corriente durante un número fijo de pasos de simulación. Al concluir dicho intervalo se reasigna una nueva referencia, y el entorno se reinicia bajo las mismas condiciones temporales para iniciar un nuevo episodio. Dentro de cada paso, el agente ejecuta el procedimiento descrito en el Algoritmo 2.2, lo que incluye la selección de acciones mediante la política de comportamiento, el almacenamiento de transiciones en el *replay buffer* y la actualización del aproximador neuronal.

Así, mientras Q-learning aprende directamente de la experiencia continua del sistema, DQN estructura el proceso en episodios consecutivos que permiten un entrenamiento más estable mediante muestreo y actualización diferenciada de las transiciones almacenadas.

3.2.5. Consideraciones adicionales

Antes de abordar los casos de estudio particulares, es necesario establecer ciertos parámetros globales del sistema eléctrico y del proceso de control que guardan relación con los entrenamientos. Estas consideraciones permiten unificar la formulación presentada anteriormente y garantizan la coherencia entre los distintos escenarios de entrenamiento y evaluación.

En primer lugar, el tiempo de muestreo constituye un elemento central tanto para la dinámica del sistema como para la interacción del agente con el entorno. En este trabajo se emplea un período

de muestreo mínimo $T_s = 10\,\mu$s (f_s = 100kHz), lo suficientemente pequeño como para capturar adecuadamente la evolución de la corriente en el inversor y respetar las restricciones de estabilidad asociadas al modelo discreto. Todas las transiciones del MDP y las actualizaciones de estado se realizan a este ritmo, de modo que el agente observa el sistema a una frecuencia constante durante el entrenamiento y la evaluación. También se evaluará cómo impacta el aumento del tiempo de muestreo en el entrenamiento y control de los agentes.

De manera adicional, los parámetros eléctricos fundamentales, como la tensión de la fuente continua, los valores de resistencia e inductancia de las cargas y las características del inversor, permanecen invariables entre los distintos casos. Esto permite que las diferencias observadas en el desempeño de los agentes se atribuyan únicamente a los algoritmos de aprendizaje y no a modificaciones en el sistema físico.

3.3. Control de corriente monofásico con agente Q-learning

En este caso, el objetivo del agente Q-learning consiste en controlar la corriente entregada por el inversor de manera que siga distintos tipos de señales de referencia. Para ello, el agente selecciona en cada instante uno de los niveles de tensión disponibles en el inversor, determinando la conmutación de los interruptores según la corriente requerida para la carga.

El caso monofásico sirve además como punto de partida para los desarrollos posteriores, ya que permite evaluar el desempeño del método tabular en un escenario de baja dimensionalidad y establecer criterios iniciales para la elección de parámetros de aprendizaje. Los resultados obtenidos en esta configuración facilitan la transición hacia los sistemas trifásicos, donde la complejidad del entorno y el tamaño de los espacios de estados y acciones aumentan significativamente.

3.3.1. Plataforma de simulación

El control basado en Q-learning para el caso monofásico se implementa sobre una plataforma de simulación integrada entre *Simulink* y *PLECS Blockset*. En esta configuración, *PLECS* se encarga de modelar el inversor de dos niveles, la fuente de continua y la carga RL (Figura 2.1), resolviendo la dinámica eléctrica del sistema con el paso de simulación fijado en el Capítulo 3. Por su parte, *Simulink* actúa como entorno de coordinación, incorporando bloques adicionales para la generación de la referencia de corriente, la adquisición de señales y la implementación del agente Q-learning.

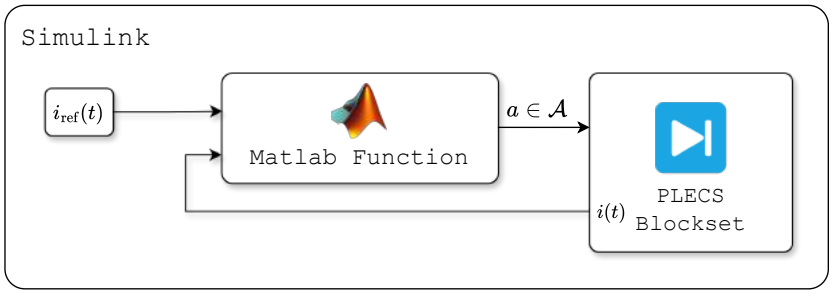


Figura 3.2: Esquema general de la plataforma de simulación para el caso monofásico con agente Q-learning.

La Figura 3.2 ilustra la estructura general empleada. El modelo eléctrico, descrito en la sección 3.1 se encuentra contenido en un subsistema de *PLECS*, desde el cual se obtiene la corriente medida en cada paso temporal. Esta señal se envía a un bloque *MATLAB Function* que implementa la

lógica del agente Q-learning y determina la acción de control a aplicar. La acción seleccionada se traduce posteriormente en señales de disparo para los interruptores del inversor, completando así el ciclo de interacción agente-sistema.

El esquema mostrado es coherente con la formulación del Proceso de Decisión de Markov presentada en la sección 3.2, ya que identifica explícitamente los elementos fundamentales de la interacción agente-entorno. La corriente de referencia $i_{\text{ref}}(t)$ y la corriente medida $i(t)$ permiten al agente determinar el estado del sistema; la salida del bloque *MATLAB Function* corresponde a la acción seleccionada en cada instante; y el modelo eléctrico en *PLECS Blockset* representa la dinámica del entorno que, tras aplicar la acción, genera el siguiente estado. De esta manera, el diagrama refleja de forma directa la estructura del MDP, permitiendo visualizar cómo se implementa físicamente el ciclo de decisión del agente dentro de la plataforma de simulación.

Esta plataforma resulta adecuada para el método tabular, dado que permite ejecutar el aprendizaje de manera totalmente *online*, actualizando la función $Q(s, a)$ durante la simulación sin necesidad de un entorno de entrenamiento separado. Además, la integración entre *Simulink* y *PLECS* facilita la realización de barridos de parámetros y el registro de métricas, los cuales se describen en capítulos posteriores.

3.3.2. Representación del agente Q-learning en la simulación

La implementación del agente Q-learning se realiza mediante un bloque *MATLAB Function* incorporado directamente en el diagrama de *Simulink*. Este bloque constituye el núcleo del proceso de toma de decisiones del agente, recibiendo en cada paso temporal la corriente medida $i(t)$ y la corriente de referencia $i_{\text{ref}}(t)$, a partir de las cuales calcula el error instantáneo y determina el estado discreto correspondiente, según la discretización definida en el presente apartado.

Una vez identificado el estado actual, el agente selecciona una acción del conjunto disponible, $a \in \{(1, 0), (0, 0), (0, 1)\}$, empleando una política ε-greedy que equilibra exploración y explotación. La acción seleccionada se traduce en las señales de disparo (S_1, S_2) aplicadas al inversor modelado en *PLECS Blockset*, generando la dinámica eléctrica que determina la siguiente observación.

De manera simultánea, dentro del mismo bloque *MATLAB Function* se actualiza la tabla de acción-valor $Q(s, a)$ en cada paso de simulación, siguiendo el esquema de aprendizaje descrito en la Algoritmo 2.1 en el Capítulo 2. Este mecanismo permite que el proceso de aprendizaje ocurra completamente *online*, sin necesidad de almacenar episodios ni utilizar memoria de repetición. La estructura de la simulación refleja directamente la organización del Proceso de Decisión de Markov (MDP), donde el bloque *MATLAB Function* opera como agente y el modelo eléctrico implementado en *PLECS Blockset* constituye el entorno dinámico. Esta correspondencia se muestra de forma explícita en el esquema de la Figura 3.2, el cual representa las señales de observación, acción y retroalimentación asociadas al ciclo de interacción agente-entorno.

En algunos casos fue necesario incorporar mecanismos adicionales de estabilización de la política, tales como *Action Hold*, *Action Sticky* y el umbral Q_{thresh}. Estas técnicas, descritas en detalle en el Apéndice A, no forman parte del algoritmo Q-learning estándar, pero resultaron relevantes para obtener un comportamiento más estable en condiciones de operación particularmente exigentes, especialmente frente a cambios bruscos de referencia o regiones del espacio de estados donde el aprendizaje tiende a ser más sensible.

Por otro lado, tal como se estableció en la sección 3.2.1 de la sección 3.2, el estado utilizado por el agente Q-learning se obtiene a partir de una representación discretizada del error instantáneo

entre la corriente referencia y la corriente en la carga,

$$s[k] = f_{\text{disc}}(i_{\text{ref}}(t) - i(t)). \tag{3.7}$$

Este error se calcula dentro de un rango acotado que cubre gran rango de la dinámica del sistema. Considerando el sistema eléctrico a controlar (apartado 3.1), se utilizará un intervalo para el error de:

$$e_{\min} = -40 \text{ A}, \qquad e_{\max} = 40 \text{ A}, \tag{3.8}$$

que permite abarcar un gran rango estados posibles.

La resolución de la discretización del estado viene determinada por el valor mínimo de error e_{step}, de modo que el número total de estados queda determinado por

$$N_e = 1 + \frac{e_{\text{máx}} - e_{\text{mín}}}{e_{\text{step}}}. \tag{3.9}$$

Cada valor continuo del error se asigna a un estado discreto mediante

$$s[k] = f_{\text{disc}}(e(t)) = \min\left(\max\left(1 + \text{floor}\left(\frac{e(t) - e_{\text{mín}}}{e_{\text{step}}}\right), 1\right), N_e\right), \tag{3.10}$$

en donde $s[k] \in \{1, 2, \ldots, N_e\}$, $\min(\cdot)$ calcula el mínimo entre dos valores, $\max(\cdot)$ calcula el máximo entre dos valores y $\text{floor}(\cdot)$ aproxima al entero menor más cercano. La Ecuación 3.10 asegura entradas discretas y acotadas en función de los valores $e_{\min}, e_{\max}$ y e_{step}

Debe destacarse que la resolución de la discretización, dada por e_{step}, influye directamente en el tamaño del espacio de estados y, en consecuencia, en la dimensión de la tabla $Q(s, a)$. Una discretización más fina (menor e_{step}) incrementa N_e y genera una tabla de mayor tamaño, en cambio, una discretización más gruesa reduce el número de estados, pero a costa de una menor resolución en la representación del error.

Con el objetivo de evaluar el impacto de la resolución del estado sobre el desempeño del controlador, en este trabajo se consideran los siguientes valores:

$$e_{\text{step}} \in \{0.1, 0.2, 0.5\} \text{ A}, \tag{3.11}$$

Considerando los valores de $e_{\text{mín}}$ y $e_{\text{máx}}$ los valores resultantes para el número de estados N_e son:

$$e_{\text{step}} = 0.1 \text{ A} \quad \Rightarrow \quad N_e = 801, \tag{3.12}$$

$$e_{\text{step}} = 0.2 \text{ A} \quad \Rightarrow \quad N_e = 401, \tag{3.13}$$

$$e_{\text{step}} = 0.5 \text{ A} \quad \Rightarrow \quad N_e = 161. \tag{3.14}$$

3.3.3. Selección de parámetros de Q-learning

El desempeño del agente Q-learning depende de manera directa de los valores asignados a sus parámetros fundamentales, los cuales determinan la forma y velocidad del aprendizaje. Con el fin de caracterizar sistemáticamente la influencia de estos parámetros sobre la calidad del control de corriente, se llevó a cabo un barrido exhaustivo que considera diversas combinaciones de la tasa de aprendizaje, el factor de descuento, la probabilidad de exploración y la resolución de la discretización del estado.

Los parámetros explorados en este estudio son los hiperparámetros propios del algoritmo Q-learning, que gobiernan la actualización de la función de acción-valor, y parámetros asociados al entorno discretizado, los cuales definen el tamaño del espacio de estados. La Tabla 3.2 resume los valores considerados para cada parámetro.

Tabla 3.2: Parámetros explorados en Q-learning

Parámetro	Valores considerados
Tasa de aprendizaje α	0.1, 0.2, 0.5, 0.6, 0.7, 0.8, 0.9
Factor de descuento γ	0.1, 0.2, 0.3, 0.4, 0.5, 0.8, 0.9
Probabilidad de exploración ε	0.01, 0.05, 0.1
Paso de discretización del error e_{step}	0.1, 0.2, 0.5

Cada combinación de parámetros genera un agente distinto que es evaluado bajo las mismas condiciones iniciales y con el mismo sistema eléctrico, lo que permite comparar de forma consistente el efecto aislado de cada parámetro sobre la calidad del seguimiento de corriente.

El número total de combinaciones exploradas es $N_{\text{comb}} = 7 \times 7 \times 3 \times 3 = 441$, lo que asegura una cobertura amplia del espacio de búsqueda. El análisis detallado del desempeño de estos agentes y la selección del conjunto óptimo de parámetros se presenta en el Capítulo 5.

3.3.4. Diseño del experimento de evaluación de parámetros

Con el fin de determinar la combinación óptima de hiperparámetros para el agente Q-learning, se diseñó un experimento sistemático basado en la simulación de un conjunto de señales de referencia que permiten evaluar el comportamiento del controlador frente a dinámicas suaves y cambios abruptos. Cada conjunto de parámetros listado en la Tabla 3.2 fue evaluado mediante una simulación independiente bajo condiciones idénticas, de manera que las diferencias observadas puedan atribuirse exclusivamente al efecto de los hiperparámetros.

El tiempo total de simulación se fijó en $T_{\text{sim}} = 4\,\text{s}$, durante el cual se aplicó al agente una secuencia de referencias de corriente dividida en cuatro intervalos consecutivos de un segundo cada uno. La secuencia se definió del siguiente modo:

1. En el primer segundo, una señal senoidal de frecuencia $50\ \text{Hz}$ y amplitud $10\ \text{A}$.
2. En el segundo segundo, la misma señal senoidal con amplitud incrementada a $25\ \text{A}$, con el fin de evaluar la capacidad del agente para adaptarse a cambios en el régimen de operación.
3. En el tercer segundo, una señal cuadrada de $50\ \text{Hz}$ y amplitud $10\ \text{A}$, introduciendo variaciones abruptas en la referencia.
4. En el cuarto segundo, una señal cuadrada de $50\ \text{Hz}$ y amplitud $20\ \text{A}$, lo que permite observar simultáneamente el efecto del cambio brusco y del aumento de nivel.

La inclusión de ambas formas de onda responde a criterios de evaluación complementarios: las señales senoidales permiten estudiar el seguimiento en regímenes suaves y periódicos, mientras que las señales cuadradas introducen discontinuidades que revelan la robustez del controlador y su capacidad de reaccionar ante transiciones rápidas.

Para todas las combinaciones de parámetros, el entrenamiento se realizó utilizando la misma función de recompensa, definida como

$$r = -\big(i_{\mathrm{ref}}(t) - i(t)\big)^2 = -e^2, \tag{3.15}$$

de modo que la selección de los parámetros del agente se base exclusivamente en el comportamiento dinámico del controlador y no en diferencias asociadas a la formulación del criterio de aprendizaje. Cabe destacar que la recompensa se calcula empleando el valor continuo del error, sin discretización.

Para cada experimento se registró la corriente medida durante todo el intervalo de simulación. Sobre dicha señal se calculó una métrica global de desempeño basada en el error acumulado a lo largo del tiempo, lo que permite ordenar las distintas combinaciones de parámetros desde el mejor hasta el peor comportamiento. Esta métrica se presenta en la sección 5.1 del Capítulo 5, donde se analizan los resultados obtenidos y se selecciona el conjunto de hiperparámetros que proporciona el desempeño más favorable.

3.4. Control de corriente monofásico con agente DQN

El enfoque basado en Q-learning permitió implementar un controlador de corriente monofásico de manera directa, utilizando una tabla de acción–valor y un espacio de estados discretizado. Este esquema resultó adecuado para el caso monofásico debido a la reducida dimensionalidad del estado, permitiendo un aprendizaje en línea dentro del entorno *Simulink + PLECS*. Sin embargo, la necesidad de discretizar explícitamente el error introduce limitaciones intrínsecas: la resolución del estado afecta el tamaño de la tabla $Q(s, a)$ y, por tanto, la capacidad del agente para capturar con precisión la dinámica del sistema.

En contraste, el enfoque de *Deep Q-Network* (DQN) reemplaza la función tabular por un aproximador neuronal capaz de operar sobre estados continuos. Esto elimina la necesidad de discretización del error y permite representar de forma más rica las relaciones no lineales presentes en el sistema eléctrico. Además, DQN incorpora elementos adicionales del aprendizaje por refuerzo moderno, como la memoria de repetición y el uso de una red objetivo, que estabilizan el proceso de aprendizaje frente a dinámicas rápidas o ruidosas.

El objetivo de esta sección es describir la configuración del agente DQN para el caso monofásico, detallando la formulación del estado continuo, la arquitectura utilizada, la estructura del entorno de entrenamiento y los parámetros involucrados en el proceso. De esta manera, se establece un marco de comparación directo con el esquema tabular presentado en la Sección 3.3, permitiendo analizar posteriormente las diferencias en desempeño entre ambos enfoques.

3.4.1. Plataforma de entrenamiento

A diferencia del esquema tabular presentado en la Sección 3.3, el agente DQN se realiza en un entorno de entrenamiento completamente programado, independiente del modelo eléctrico implementado en *Simulink*. En este trabajo, tanto el agente como el entorno se desarrollan en *MATLAB* utilizando las herramientas del *Reinforcement Learning Toolbox*, lo que permite operar directamente sobre estados continuos y emplear redes neuronales profundas como aproximador de la función Q.

La Figura 3.3 muestra la estructura general de la plataforma de entrenamiento. El agente DQN, configurado con la arquitectura y los hiperparámetros correspondientes, interactúa con el entorno a través de dos componentes fundamentales: la función *step*, encargada de actualizar el estado en

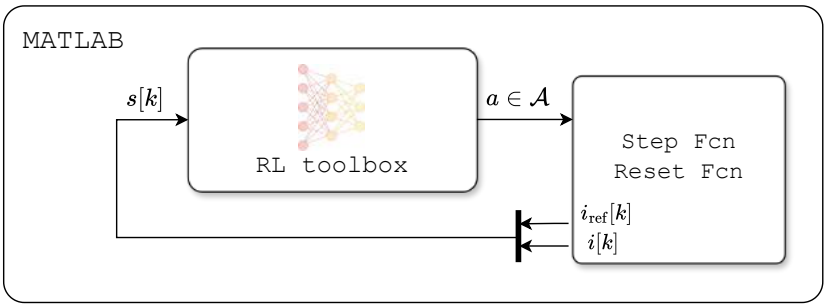


Figura 3.3: Esquema general de la plataforma de entrenamiento para el caso monofásico con agente DQN.

cada paso del episodio, y la función *reset*, que establece las condiciones iniciales al comienzo de cada nuevo episodio. Ambas funciones implementan de manera explícita la dinámica discreta del inversor monofásico y de la carga RL, utilizando el mismo período de muestreo empleado en la simulación física.

En cada paso de entrenamiento, el agente observa el estado continuo generado por la función *step*, selecciona una acción mediante la política de comportamiento y recibe su recompensa asociada. La transición resultante (s_k, a_k, r_k, s_{k+1}) se almacena en la memoria de repetición del agente, desde donde se extraen *minibatches* para actualizar la red neuronal y aproximar la función $Q(s, a)$ de manera estable. Esta interacción entre agente y entorno se repite durante un conjunto de episodios hasta completar el proceso de aprendizaje. De manera estándar el entrenamiento del agente sigue el Algoritmo 2.2, mostrado en el Capítulo 2.

3.4.2. Representación del agente DQN en la simulación

El agente DQN utilizado en este trabajo se implementa mediante las herramientas del *Reinforcement Learning Toolbox* de *MATLAB* [22], siguiendo la estructura estándar del algoritmo. A continuación se detalla toda la configuración del agente para realizar el entrenamiento con la herramienta recién mencionada.

3.4.2.a. Espacio de observaciones y acciones

- **Observaciones:** Vector $s[k]$ de `rlNumericSpec` para observaciones de $i_{\text{ref}}(t)$ e $i(t)$.
- **Acciones:** Conjunto finito de 3 acciones `rlFiniteSetSpec`.

3.4.2.b. Configuración de la Q-Network

Corresponde a la estructura fundamental de la red neuronal. Esquemáticamente corresponde a la misma arquitectura mostrada en la Figura 2.5 del Capítulo 2. La red objetivo ($\hat{Q}$-Network) es idéntica a la red principal.

- **Entrada:** Vector de observación.
- **Capas ocultas:** 2 capas totalmente conectadas con N_n neuronas y función de activación *ReLU*, cada una.

- **Salida:** Capa totalmente conectada con tres neuronas que representan los valores estimados de $Q(s, a)$ para cada acción.

3.4.2.c. Opciones de entrenamiento del agente

Entre las opciones disponibles en la *RL Toolbox* para personalizar el agente durante su entrenamiento, se incluyen los distintos elementos descritos en la sección 2.5.2 del Capítulo 2:

- **Double DQN**: se hará uso de esta opción `'UseDoubleDQN'=true`.
- **Actualización de red objetivo:** `'TargetUpdateMethod'='smoothing'`
- **Factor de actualización de objetivo (TSF):** `'TargetSmoothFactor'= 1e-3`.
- **Tamaño Memoria de Experiencia:** Cantidad de transiciones (s_k, a_k, r_k, s_{k+1}) almacenadas en memoria. `'ExperienceBufferLength'=1e5`.
- **Exploración:** Política de ε-greedy con valor inicial $\varepsilon_{\text{ini}} = 1$ y valor final $\varepsilon_{\text{min}} = 0{,}01$.

El resto de los parámetros ajustables, como la *tasa de aprendizaje*, el *factor de descuento*, el *tamaño del minibatch*, entre otros, se analizarán mediante un barrido de parámetros detallados en la sección 3.4.3.

Un componente importante para el aprendizaje es la exploración del agente durante el entrenamiento. Se rige mediante la Ecuación 2.29 (Capítulo 2), que decae en cada paso de tiempo, por que el valor de $\varepsilon_{\text{decay}}$ depende de qué tanta exploración se desea realizar. Dado que la exploración decae gradualmente desde un valor inicial a su valor mínimo, es necesario determinar previamente en qué punto del entrenamiento debe llegar al valor mínimo. Si un entrenamiento de E episodios, en donde cada episodio tiene N_{step} pasos de simulación. Para determinar el $\varepsilon_{\text{decay}}$ que permita decaer la exploración al mínimo en el episodio E_ε viene dado por:

$$\varepsilon_{\text{decay}} = 1 - \sqrt[E_\varepsilon \, N_{\text{step}}]{\left(\frac{\varepsilon_{\text{mín}}}{\varepsilon_{\text{ini}}}\right)}, \tag{3.16}$$

En este trabajo se evalúa el efecto de hacer decaer la exploración hasta su valor mínimo en distintos instantes del entrenamiento, definidos como $E_\varepsilon = E,\ E/2,\ E/3$, donde E corresponde al número total de episodios. Esto permite analizar el impacto que tiene la rapidez del decaimiento en el proceso de aprendizaje.

3.4.3. Selección de parámetros de Deep Q-Network

Al igual que en el método tabular, el desempeño del agente DQN depende de los valores asignados a sus parámetros de configuración. Sin embargo, en este caso el proceso de aprendizaje incorpora elementos adicionales, como la arquitectura de la red neuronal, el tamaño de los *mini-batches*, el *target smoothing factor* y el mecanismo de *experience replay*, que amplían significativamente el espacio de búsqueda. Por este motivo, resulta indispensable explorar distintas combinaciones de hiperparámetros para determinar qué configuraciones conducen a un aprendizaje estable y a un seguimiento adecuado de la referencia.

Tabla 3.3: Parámetros considerados en la exploración de DQN.

Parámetro	Valores
Neuronas por capa N_n	24, 32, 64
Tasa de aprendizaje α	5e-4, 1e-3, 5e-3
Factor de descuento γ	0.95, 0.97, 0.99
Decaimiento de exploración $\varepsilon_{\text{decay}}$	Ecuación 3.16
Target Smooth Factor TSF	1e-3, 5e-4, 1e-4
Tamaño de miniBatch $\lvert\mathcal{M}\rvert$	64, 128, 256

Un aspecto relevante es que el decaimiento de exploración $\varepsilon_{\text{decay}}$ no se fijó como un valor único, sino que se calculó de manera específica para cada configuración a partir de la Ecuación 3.16, dependiendo del número de episodios y número total de pasos. Esto garantiza que el agente mantenga un equilibrio adecuado entre exploración y explotación en función de la duración efectiva del entrenamiento.

Aunque la Tabla 3.3 resume todos los parámetros relevantes, no todas las combinaciones fueron evaluadas en cada etapa del proceso. El tiempo de entrenamiento depende directamente del número de pasos por episodio, del tiempo de muestreo y de la complejidad de la red neuronal. Por ejemplo, una sola ejecución con 24 combinaciones puede requerir del orden de una hora de simulación. Por esta razón, se empleó una estrategia iterativa: primero se realizó un barrido amplio para identificar rangos prometedores y, posteriormente, se acotaron los hiperparámetros para efectuar exploraciones más focalizadas.

Considerando el conjunto de valores posibles, el número total de combinaciones asciende a $N_{\text{comb}} = 3 \times 3 \times 3 \times 3 \times 3 \times 3 = 729$, lo que proporciona una cobertura amplia del espacio de hiperparámetros. Si además se consideran los análisis para diferentes función de recompensas o tiempos de muestreo, dicho número podría incrementar significativamente. El análisis detallado de los resultados obtenidos y la selección del agente óptimo se presentan en el Capítulo 5.

3.4.4. Diseño del experimento de evaluación de parámetros

El objetivo de esta etapa fue determinar configuraciones adecuadas de hiperparámetros y condiciones de entrenamiento para el agente DQN aplicado al control de corriente monofásico. Dado que el impacto de cada parámetro sobre el comportamiento del agente no era evidente a priori, se adoptó una estrategia experimental de carácter iterativo, estructurada en fases sucesivas que permitieron acotar de manera progresiva el espacio de búsqueda. Otro aspecto relevante es la definición del entorno (funciones *Step* y *Reset*), ya que, al estar completamente programado en *scripts* de MATLAB, requiere modelar la dinámica del sistema eléctrico, además de generar la señal de referencia a partir de la cual el agente aprende la política deseada.

En primera instancia, la función *Step* se ejecuta en cada paso $T_{\text{s}} = 10\mu\text{s}$, por lo que la dinámica del sistema eléctrico se aproxima por diferencias finitas (Apéndice C). En el Algoritmo 3.1 se muestra el procedimiento que sigue la función *Step/Reset* durante el entrenamiento.

Algoritmo 3.1 Interacción agente-entorno mediante funciones *Reset* y *Step*

para cada *episodio* **hacer**
  Función Reset:
  Inicializar el tiempo $t \leftarrow 0$
  Inicializar la corriente i_0 y la amplitud A (por ejemplo, $A \sim \mathcal{U}[1, 20]$)
  Calcular la referencia inicial $i_{\text{ref}}(0) = A \sin(\omega t)$
  Definir el estado inicial del entorno: $S_0 = [\, i_0,\ i_{\text{ref}}(0)\,]$
  para cada paso de tiempo $t = 0, \ldots, N_{\text{step}} - 1$ **hacer**
    El agente selecciona una acción a_t según la política aprendida o una estrategia ε-greedy
    Mapear la acción a_t a la tensión aplicada: $V_t \in \{-V_{\text{dc}},\ 0,\ +V_{\text{dc}}\}$
    Actualizar la corriente mediante

$$i_{t+1} = i_t + T_{\text{s}} \frac{V_t - R\, i_t}{L}$$

    Actualizar el tiempo: $t \leftarrow t + T_{\text{s}}$
    Calcular la referencia senoidal en el nuevo instante:

$$i_{\text{ref}}(t) = A \sin(\omega t)$$

    Calcular el error normalizado:

$$e_N = \frac{i_{\text{ref}}(t) - i_{t+1}}{A}$$

    Calcular la recompensa inmediata R_t:

$$R_t = 1 - e_N^2$$

    Construir el nuevo estado:

$$S_{t+1} = [\, i_{t+1},\ i_{\text{ref}}(t)\,]$$

    si $t \geq T_{\text{episode}}$ **entonces**
      Terminar episodio
      break

3.4.4.a. Fase I: Exploración Amplia del Espacio de Hiperparámetros

En la primera etapa se realizó un barrido general con el fin de identificar regiones prometedoras del espacio de hiperparámetros. Para ello se definieron condiciones iniciales exigentes:

- Tiempo de muestreo fijo de $T_s = 1 \times 10^{-5}\,\text{s}$, correspondiente a una frecuencia de muestreo de $f_s = 100\,\text{kHz}$.
- Duración del episodio igual a $0{,}04\,\text{s}$, lo que abarca aproximadamente dos ciclos completos de la señal de referencia.
- Entrenamientos de 300 episodios por agente.
- Referencia senoidal de $50\,\text{Hz}$ cuya amplitud, para cada episodio, se seleccionó de forma aleatoria según $A = \text{randi}([1, 20])$.

El barrido incluyó variaciones de hiperparámetros tales como el *learning rate*, el factor de descuento, el decaimiento del parámetro ε, la arquitectura de la red neuronal y el tamaño del *experience buffer*, entre otros. Para cada configuración entrenada se registró la evolución de la recompensa promedio utilizando el *training progress* del *Reinforcement Learning Toolbox*.

Los resultados de esta fase evidenciaron comportamientos de aprendizaje inestables en un número significativo de configuraciones, manifestados como oscilaciones abruptas en las recompensas promedio. Esto sugirió una alta sensibilidad del agente tanto a los hiperparámetros seleccionados como a la resolución temporal establecida por el tiempo de muestreo.

3.4.4.b. Fase II: Ajustes de la Dinámica del Entorno y Estabilidad del Aprendizaje

A partir de las observaciones obtenidas en la fase anterior se llevó a cabo una segunda etapa cuyo objetivo fue evaluar el impacto del tiempo de muestreo y del número de episodios sobre la estabilidad del aprendizaje del agente.

- **Modificación del Tiempo de Muestreo:**

  Se realizaron nuevos entrenamientos utilizando un tiempo de muestreo mayor:

  $$T_s = 2{,}5 \times 10^{-5}\,\mathrm{s} \quad (f_s = 40\,\mathrm{kHz}),$$

  junto con un aumento del número de episodios a 1000. El propósito era determinar si una dinámica menos exigente desde el punto de vista temporal facilitaba el aprendizaje del agente. Los resultados mostraron que el agente lograba converger hacia buenas recompensas alrededor del episodio 200, pero la estabilidad no mejoraba por extender el entrenamiento sin modificar otros parámetros: la recompensa tendía a volverse errática en etapas avanzadas, posiblemente debido a la sobrescritura del *experience buffer* y a un decaimiento de ε demasiado agresivo.

- **Ajuste de decaimiento de exploración:**

  Para evitar una transición prematura desde la exploración hacia la explotación, se evaluaron distintas estrategias de decaimiento del parámetro ε, ajustándolo para que el decaimiento completo coincidiera con distintas proporciones del entrenamiento: a la totalidad del episodio, a la mitad y al primer tercio. Este ajuste permitió lograr una exploración más equilibrada y un aprendizaje más estable en comparación con los barridos iniciales.

3.4.4.c. Fase III: Análisis Univariable y Evaluación de Funciones de Recompensa

Tras las conclusiones obtenidas en las dos etapas previas, se diseñó una tercera fase orientada a estudiar de manera aislada el efecto de cada hiperparámetro sobre el desempeño del agente. El objetivo de esta etapa fue acotar de forma sistemática las combinaciones más prometedoras, reduciendo la variabilidad observada en los barridos anteriores y delimitando así una región de búsqueda más estable para fases posteriores de optimización.

Para ello, se consolidaron en un único conjunto los resultados de todos los entrenamientos realizados, registrando métricas tales como el RMSE, la recompensa promedio y la estabilidad del aprendizaje. Con esta información se efectuó un análisis univariable, examinando para cada hiperparámetro la distribución del desempeño de los agentes cuando dicho parámetro varía de

manera individual, manteniendo los demás dentro de rangos razonables. Este procedimiento permitió identificar tendencias claras en parámetros como el *learning rate*, el factor de descuento γ, el tamaño del *minibatch* y el decaimiento de ε, delimitando rangos en los que el aprendizaje resulta más estable y eficiente.

Adicionalmente, se llevó a cabo un experimento orientado a comparar distintas funciones de recompensa con el fin de evaluar su impacto en la estabilidad del entrenamiento y la capacidad de seguimiento de corriente. Se analizaron formulaciones cuadráticas, penalizaciones basadas en bandas, términos exponenciales y funciones que incorporan el costo por conmutación. Este estudio permitió identificar qué estructuras de recompensa favorecen un comportamiento más consistente del agente y contribuyen a una mejor generalización.

3.5. Configuración de los agentes trifásicos de control de corriente

En esta sección se describe la configuración de los agentes de aprendizaje por refuerzo utilizados para el control de corriente trifásico, considerando tanto el enfoque tabular basado en Q-learning como el enfoque basado en Deep Q-Network (DQN). Para cada algoritmo se analizan dos marcos de referencia distintos, $\alpha\beta$ y dq, detallando las variables de observación, el espacio de acciones y las decisiones de diseño adoptadas en cada caso.

3.5.1. Agente Q-learning trifásico en marco $\alpha\beta$

En el caso del control de corriente trifásico en el marco de referencia $\alpha\beta$, el agente Q-learning utiliza como observaciones únicamente los errores de corriente en ambas componentes, definidos como

$$e_\alpha = i_{\alpha,\mathrm{ref}} - i_\alpha, \qquad e_\beta = i_{\beta,\mathrm{ref}} - i_\beta. \tag{3.17}$$

Estas variables son discretizadas de manera uniforme, construyendo un espacio de estados bidimensional que permite representar la dinámica del error de corriente en el plano estacionario.

El espacio de acciones del agente está conformado por los ocho estados de conmutación posibles del inversor trifásico de dos niveles. Cada acción corresponde a un vector de tensión aplicado al sistema, representado implícitamente a través de las señales de disparo de los interruptores del inversor. De este modo, el agente aprende una política que asocia pares discretizados de errores (e_α, e_β) con decisiones de conmutación específicas.

Esta configuración permite al agente aprender de manera *online* una política de control directamente en el marco $\alpha\beta$, sin requerir transformaciones adicionales ni información explícita del ángulo eléctrico.

3.5.2. Agente Q-learning trifásico en marco dq

Para el control trifásico en el marco de referencia dq, el agente Q-learning emplea como observaciones únicamente los errores de corriente en las componentes directa y en cuadratura, definidos como

$$e_d = i_{d,\mathrm{ref}} - i_d, \qquad e_q = i_{q,\mathrm{ref}} - i_q. \tag{3.18}$$

Al igual que en el caso $\alpha\beta$, estas variables son discretizadas de forma uniforme, dando lugar a un espacio de estados bidimensional.

El conjunto de acciones disponible para el agente es idéntico al utilizado en el marco $\alpha\beta$, correspondiendo a los ocho estados de conmutación del inversor trifásico. De esta forma, la diferencia entre ambos enfoques radica exclusivamente en la representación del estado, manteniendo inalterado el espacio de acciones.

Esta formulación permite evaluar el desempeño del aprendizaje por refuerzo en un marco de referencia rotante, alineado con la estructura clásica de control FOC, sin introducir modificaciones adicionales en la lógica de decisión del agente.

3.5.3. Agente DQN trifásico en marco $\alpha\beta$

En el enfoque basado en Deep Q-Network, el control trifásico en el marco $\alpha\beta$ utiliza como observaciones únicamente los errores de corriente e_α y e_β, sin incorporar información adicional sobre el ángulo eléctrico o el sector del vector de tensión. La elección de trabajar con errores, en lugar de corrientes absolutas, permite reducir la dimensionalidad del espacio de estados y facilita el proceso de aprendizaje del agente.

El espacio de acciones corresponde nuevamente a los ocho estados de conmutación del inversor trifásico. A diferencia del Q-learning tabular, el agente DQN es entrenado de manera *offline* en el entorno MATLAB, utilizando una red neuronal como aproximador de la función de acción-valor. Una vez finalizado el entrenamiento, la política aprendida es exportada y utilizada en forma determinista mediante una tabla de consulta (*Look-Up Table*, LUT).

El entrenamiento del agente DQN trifásico se realiza con un período de muestreo de $T_s = 4 \times 10^{-5}\,\mathrm{s}$, coherente con las condiciones utilizadas en los casos monofásicos y compatible con la dinámica del sistema eléctrico considerado.

3.5.4. Agente DQN trifásico en marco dq

En el control trifásico basado en Deep Q-Network en el marco de referencia dq, el vector de observaciones está compuesto por los errores de corriente en las componentes directa y en cuadratura, junto con una variable discreta asociada al sector eléctrico, denotada como N_θ.

El ángulo eléctrico se divide en 12 sectores uniformes, de modo que la variable N_θ toma valores enteros en el conjunto $N_\theta \in \{0, 1, \ldots, 11\}$. La inclusión de esta información resulta necesaria debido a que el agente no logra generalizar adecuadamente su política utilizando únicamente los errores en el marco dq, como consecuencia del carácter periódico del sistema de referencia y de la simetría inherente del inversor trifásico.

El uso explícito del sector permite desambiguar la relación entre los errores de corriente y las acciones de conmutación óptimas, facilitando el aprendizaje de una política coherente a lo largo de todo el ciclo eléctrico. El espacio de acciones está conformado por los ocho estados de conmutación del inversor trifásico de dos niveles.

El entrenamiento del agente DQN trifásico en marco dq se realiza manteniendo un período de muestreo de $T_s = 4 \times 10^{-5}\,\mathrm{s}$. Al igual que en el marco $\alpha\beta$, la política aprendida es exportada como una tabla de consulta (*Look-Up Table*, LUT), permitiendo su implementación eficiente en la plataforma de simulación final.

Capítulo 4

Implementación de esquemas de control en PLECS

En los capítulos anteriores se definió la formulación del problema de control como un Proceso de Decisión de Markov y se estableció la configuración de los agentes de aprendizaje por refuerzo utilizados en este trabajo. Sobre dicha base, el presente capítulo aborda la implementación de los esquemas de control desarrollados en una plataforma de simulación orientada a la evaluación del desempeño del controlador final.

El objetivo de este capítulo es describir el proceso de homologación de los agentes entrenados en el entorno *MATLAB/Simulink* hacia *PLECS Standalone*, plataforma utilizada para analizar su comportamiento bajo condiciones de operación más cercanas a una ejecución embebida. En particular, se presentan las estrategias de implementación adoptadas para los controladores basados en Q-learning y Deep Q-Network, considerando las diferencias inherentes a cada enfoque y las restricciones computacionales asociadas a la simulación a alta frecuencia.

Para el caso del Q-learning, se detalla la implementación directa de la lógica del agente mediante un bloque *C-Script*, replicando el comportamiento del controlador tabular utilizado durante el entrenamiento. En el caso del controlador DQN, se introduce una estrategia alternativa basada en la exportación de la política aprendida como una tabla de consulta (*Look-Up Table*), permitiendo preservar el comportamiento del agente sin necesidad de ejecutar la red neuronal durante la simulación.

Las implementaciones descritas en este capítulo constituyen el marco común sobre el cual se desarrollan las pruebas de desempeño y robustez presentadas en capítulos posteriores, permitiendo una comparación consistente entre las distintas metodologías de control evaluadas.

4.1. Implementación del agente Q learning en *PLECS Standalone*

Una vez seleccionada la mejor combinación de parámetros, el agente Q-learning resultante se homologa en *PLECS Standalone* con el fin de evaluar su desempeño en la plataforma final del controlador. Esta etapa es necesaria ya que el proceso de entrenamiento y selección se realiza en el entorno integrado *Simulink + PLECS Blockset*, mientras que la validación y el funcionamiento operativo del controlador deben ejecutarse directamente en *PLECS*, utilizando un bloque *C-Script* que replica la lógica del agente.

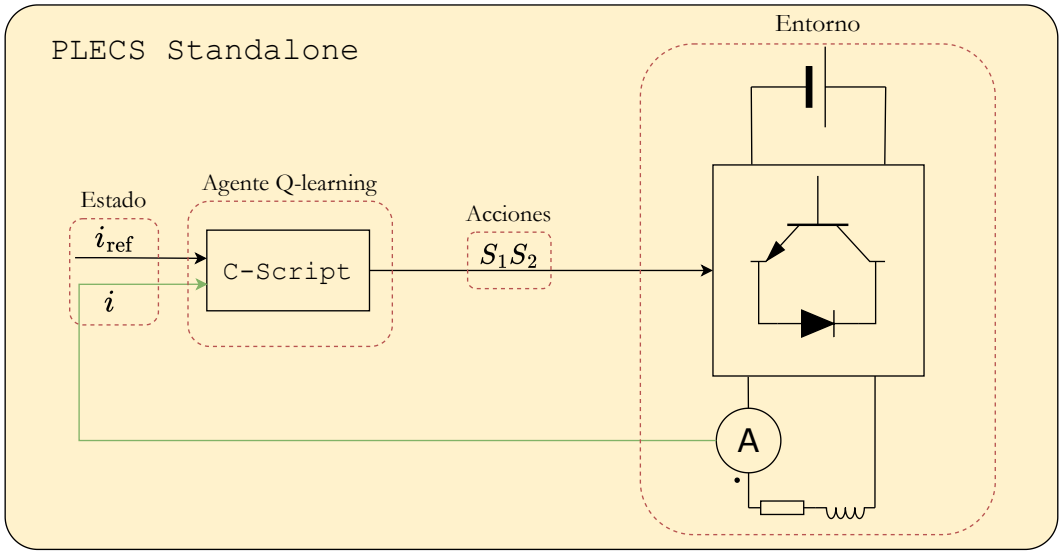


Figura 4.1: Control de corriente con agente Q-learning en PLECS Standalone.

En la Figura 4.1 se muestra el esquema de control dentro de PLECS Standalone, en donde *C-Script* implementa la misma lógica que el bloque *MATLAB Function* utilizado en el entorno *Simulink* (Figura 3.2). Este bloque recibe como entradas la corriente medida y la corriente de referencia, a partir de las cuales calcula la discretización del error, calcula la recompensa, aplica la política ε-greedy y evalúa la función de acción-valor $Q(s, a)$. La salida del *C-Script* corresponde a las señales de disparo que se aplican a los IGBT del inversor.

La implementación en *PLECS Standalone* permite evaluar al controlador en un entorno más cercano a una ejecución embarcada, sin depender del bloque *MATLAB Function*. Asimismo, esta configuración facilita el análisis del desempeño bajo condiciones adicionales que no forman parte del proceso de selección de hiperparámetros. Dichas evaluaciones complementarias se presentan en detalle en el Capítulo 5.

Una vez homologado el agente en la plataforma *PLECS Standalone*, se realizan pruebas adicionales destinadas a evaluar su capacidad de generalización frente a condiciones no contempladas durante el proceso de entrenamiento. En particular, se consideran ensayos con cambios de amplitud en la referencia, variaciones en la frecuencia de la señal deseada y modificaciones en los parámetros de carga, incluyendo componentes resistivas e inductivas. Estas pruebas permiten analizar la robustez del controlador ante dinámicas distintas a las utilizadas en la etapa de aprendizaje y determinar su comportamiento frente a perturbaciones que afectan de forma directa la respuesta del sistema eléctrico.

4.2. Implementación del Controlador DQN en PLECS mediante Tabla LUT

Una vez seleccionado el agente DQN con el mejor desempeño en los experimentos de entrenamiento, se procedió a su implementación en el entorno de simulación PLECS. Debido a que ejecutar una red neuronal profunda dentro de un *C-Script* en simulaciones de alta frecuencia resulta computacionalmente costoso, se adoptó una estrategia basada en la exportación de la política aprendida como una tabla de consulta (*Look-Up Table*, LUT). Este enfoque permite reproducir la política del agente de manera determinista, eficiente y con tiempos de simulación significativamente menores.

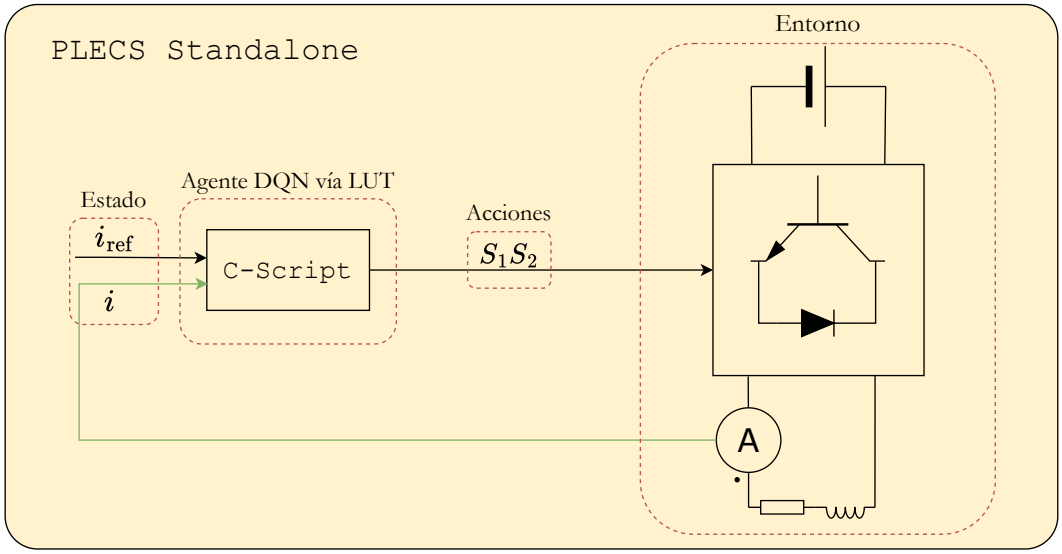


Figura 4.2: Control de corriente con agente Deep Q-Network en PLECS Standalone.

En la Figura 4.2 se puede observar que es la misma estructura que para el agente Q-learning, lo cual era esperable dado que ambos problemas son formulaciones de procesos de decisión de Markov. La LUT se genera evaluando la política del agente entrenado en modo completamente determinista (sin exploración), sobre una malla bidimensional que discretiza las dos variables observadas por el controlador: la corriente medida i y la referencia de corriente i_{ref}. Para ello se define un rango simétrico en torno a cero y una resolución ajustable, de modo que la malla pueda exportarse con diferentes niveles de granularidad según los requerimientos del sistema. Cada punto de la malla corresponde a un par (i, i_{ref}) y la tabla almacena directamente la acción que maximiza el valor Q para dicho estado, es decir:

$$\text{LUT}(i, i_{\text{ref}}) = \arg\max_{a} Q([i, i_{\text{ref}}], a), \tag{4.1}$$

Este archivo puede generarse con distintas resoluciones (por ejemplo, pasos de 0.5 A, 0.25 A o 0.1 A), lo que permite equilibrar precisión de la política y costo computacional en la simulación.

En PLECS, la LUT se incluye dentro de un `C-Script` que recibe la corriente medida y la referencia de corriente del sistema eléctrico, discretiza ambas magnitudes según los parámetros definidos y realiza una búsqueda directa en la tabla. La salida del bloque corresponde a la acción escogida por la política del agente, es decir, una de las tres decisiones posibles del inversor: $-V_{\text{dc}}$, 0, o $+V_{\text{dc}}$.

Con esta estructura, el controlador DQN puede evaluarse en simulaciones en otra plataforma sin necesidad de cargar la red neuronal completa. De esta forma, la política aprendida en MATLAB queda homologada en PLECS, preservando el comportamiento del agente a la vez que se mantiene un costo computacional compatible con estudios de dinámica del inversor.

4.3. Implementación de los controles trifásicos de corriente en PLECS

En esta sección se describe la implementación en *PLECS Standalone* de los esquemas de control de corriente trifásico basados en aprendizaje por refuerzo, considerando tanto agentes Q-learning como Deep Q-Network (DQN), y los marcos de referencia $\alpha\beta$ y dq. A diferencia de la etapa de configuración y entrenamiento, presentada en el capítulo anterior, el foco aquí se centra en la estructura de control utilizada en la plataforma final de simulación y en la forma en que los agentes se integran al esquema del inversor trifásico.

En todos los casos, el control se implementa mediante bloques *C-Script*, los cuales permiten replicar la lógica del agente de aprendizaje por refuerzo y generar directamente las señales de disparo del inversor, sin dependencia del entorno MATLAB. Esta aproximación permite evaluar el desempeño de los controladores en un entorno más cercano a una implementación embebida, manteniendo tiempos de simulación compatibles con estudios de dinámica eléctrica.

4.3.1. Estructura general del control trifásico

La estructura base del control trifásico es común a todos los casos analizados. Las corrientes de fase medidas (i_a, i_b, i_c) son transformadas al marco de referencia correspondiente mediante las transformaciones de Clarke y, cuando aplica, de Park. A partir de estas variables se calculan los errores de corriente, los cuales constituyen las entradas al agente de aprendizaje por refuerzo.

El agente procesa esta información y selecciona una acción discreta asociada a uno de los ocho estados de conmutación del inversor trifásico de dos niveles. Dicho estado se traduce directamente en las señales de disparo (S_a, S_b, S_c) aplicadas a los interruptores del inversor, cerrando el lazo de control de corriente.

4.3.2. Implementación del control trifásico Q-learning en PLECS

En el caso del control trifásico basado en Q-learning, el agente se implementa completamente dentro de un bloque *C-Script*, manteniendo una tabla Q que se actualiza de manera *online* durante la simulación. El bloque recibe como entradas los errores de corriente definidos en el marco de referencia correspondiente y ejecuta la política ε-greedy para la selección de acciones.

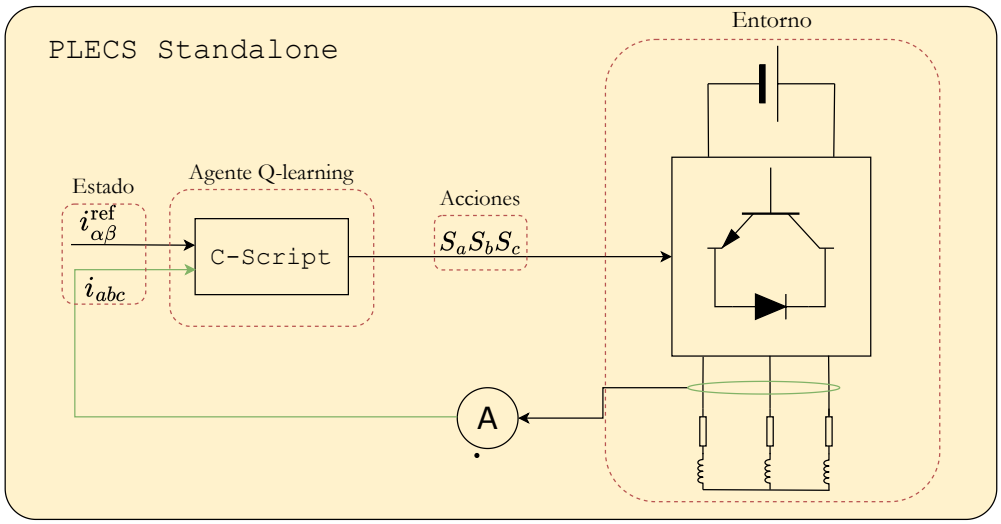


Figura 4.3: Control de corriente trifásico en marco $\alpha\beta$ con agente Q-learning en PLECS Standalone.

Como se muestra en la Figura 4.3 para marco $\alpha\beta$ y con carga RL, el *C-Script* recibe como entradas las corrientes medidas i_{abc} y las corrientes de referencia $i_{\alpha\beta}$. Internamente calcula los errores de corriente e_α y e_β y los discretiza según los parámetros definidos durante la configuración del agente.

Se puede apreciar en la Figura 4.4 que en el marco dq, el bloque *C-Script* recibe las corrientes medidas i_{abc}, las corrientes de referencia i_{dq} y además una frecuencia de referencia sobre la cual rota el marco sincrínico dq.

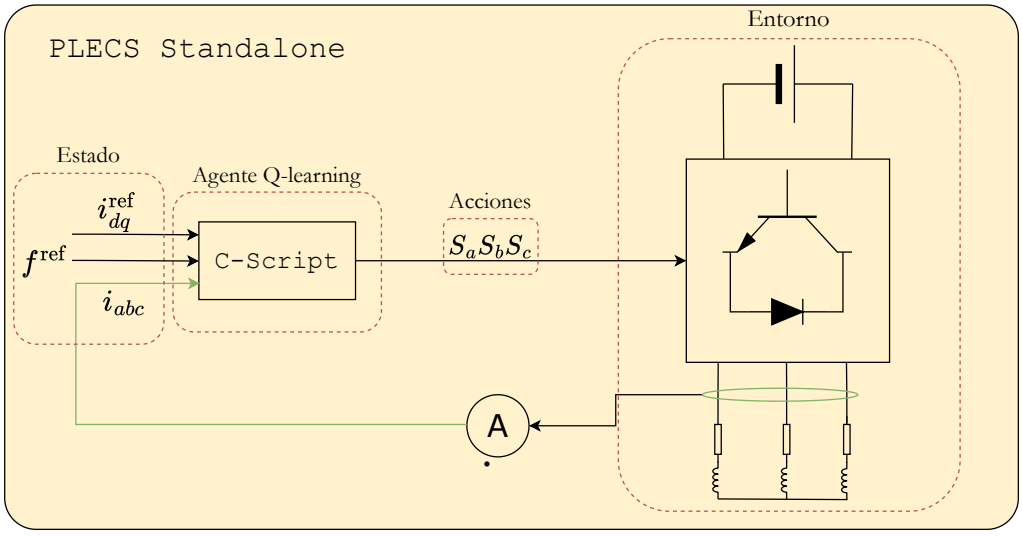


Figura 4.4: Control de corriente trifásico en marco dq con agente Q-learning en PLECS Standalone.

Este control, si bien la frecuencia de referencia es ingresada al *C-Script*, esta no ingresa como estado a la tabla Q directamente sino que lo hace a través de las transformaciones de Park, y luego al calcular el error de corriente entre para ambos ejes. Esto permite mantener una tabla de dos dimensiones evitando crecer indiscriminadamente agregando variables por discretizar.

En ambos casos, la salida del bloque corresponde a las señales de disparo del inversor, generadas a partir del estado de conmutación seleccionado. El aprendizaje se produce en tiempo de simulación, permitiendo al agente adaptar su política frente a variaciones en la carga, la referencia o las condiciones iniciales del sistema.

4.3.3. Implementación del control trifásico DQN en PLECS

Para el control trifásico basado en Deep Q-Network, la implementación en *PLECS* se realiza mediante la ejecución determinista de la política aprendida durante la etapa de entrenamiento. Dado que la evaluación directa de una red neuronal dentro de un *C-Script* resulta computacionalmente costosa, la política del agente se exporta previamente como una tabla de consulta (*Look-Up Table*, LUT).

El esquema de control implementado en PLECS para el DQN trifásico es idéntico a los mostrados en las Figuras 4.3 y Figura 4.4, para los marco $\alpha\beta$ y dq, respectivamente, la diferencia solo radica en cómo el agente realiza la elección de acciones.

En el caso trifásico con marco de referencia $\alpha\beta$, el agente internamente solo utiliza los errores e_α y e_β para la toma de decisiones. Por otro lado, en el marco dq, además de los errores e_d y e_q, se incorpora la variable discreta asociada al sector eléctrico N_θ, definida a partir de la división del ángulo eléctrico en doce sectores discretos uniformes.

A partir de estas entradas, el *C-Script* discretiza las variables según los parámetros definidos durante la exportación de la LUT y realiza una búsqueda directa para determinar la acción óptima. La acción seleccionada corresponde a uno de los ocho estados de conmutación del inversor, los cuales se traducen en las señales de disparo aplicadas al sistema de potencia.

Esta estrategia permite reproducir fielmente el comportamiento del agente DQN entrenado en MATLAB, manteniendo una carga computacional reducida y facilitando la evaluación del controlador en simulaciones de alta frecuencia.

En los casos trifásicos basados en Deep Q-Network, la política aprendida durante el entrenamiento en MATLAB se exporta a *PLECS* en forma de una tabla de consulta (*Look-Up Table*, LUT). Esta tabla se construye evaluando de manera determinista la función de acción–valor aproximada por la red neuronal, seleccionando para cada estado la acción que maximiza el valor Q.

En el caso monofásico, la LUT se genera evaluando la política sobre una malla definida por la corriente medida y la referencia de corriente, de acuerdo con la Ecuación 4.1.

Para el control trifásico en el marco de referencia $\alpha\beta$, la construcción de la LUT se extiende naturalmente al considerar como estado los errores de corriente en ambas componentes. En este caso, la política exportada queda definida como

$$\mathrm{LUT}(e_\alpha, e_\beta) = \arg\max_a Q([e_\alpha, e_\beta], a)\,, \tag{4.2}$$

donde e_α y e_β corresponden a los errores de corriente en el marco estacionario. Cada par discretizado (e_α, e_β) se asocia a una de las ocho acciones posibles del inversor trifásico de dos niveles.

En el caso del control trifásico en el marco dq, la política aprendida depende no solo de los errores de corriente, sino también del sector eléctrico asociado al ángulo del vector de tensión. En consecuencia, la LUT se construye evaluando la función Q sobre un espacio de estados

tridimensional, quedando definida como

$$\mathrm{LUT}(e_d, e_q, N_\theta) = \arg\max_a Q([e_d, e_q, N_\theta], a)\,, \tag{4.3}$$

donde e_d y e_q son los errores de corriente en el marco rotante, y N_θ es una variable discreta que identifica el sector eléctrico, considerando una división uniforme del ángulo en doce sectores.

En todos los casos, la LUT almacena directamente la acción óptima para cada combinación discretizada de las variables de estado, permitiendo reproducir de manera determinista la política aprendida por el agente DQN sin necesidad de ejecutar la red neuronal durante la simulación en *PLECS*. Esta estrategia reduce significativamente el costo computacional del control y facilita su evaluación en simulaciones de alta frecuencia.

4.3.4. Diferencias prácticas entre Q-learning y DQN en PLECS

Desde el punto de vista de la implementación, la principal diferencia entre ambos enfoques radica en la forma en que se gestiona la política de control. En Q-learning, la tabla Q se actualiza en tiempo real, lo que permite al agente adaptarse a cambios no considerados durante el diseño inicial, a costa de una mayor variabilidad transitoria. En contraste, el enfoque DQN ejecuta una política fija, previamente entrenada, lo que elimina los transitorios asociados al aprendizaje *online*, pero no puede lograr un control fuera del rango de entrenamiento dado que en este caso la política implementada como LUT es rígida en ese sentido.

Estas diferencias se reflejan en el comportamiento observado en las simulaciones en *PLECS*, las cuales se analizan en detalle en el capítulo siguiente.

Capítulo 5

Análisis de impacto de hiperparámetros en aprendizaje de agentes

En este capítulo se analiza cómo la selección de hiperparámetros afecta el proceso de aprendizaje y el desempeño final de agentes basados en aprendizaje por refuerzo aplicados al control de corriente monofásico. A partir de la configuración experimental descrita en el Capítulo 3, se evalúan de manera sistemática múltiples combinaciones de parámetros para los enfoques Q-learning y *Deep Q-Network* (DQN), con el objetivo de identificar tendencias, regiones de operación favorables y configuraciones que conduzcan a políticas de control estables.

Para cuantificar el desempeño, se definen métricas orientadas a capturar tanto la calidad del seguimiento de referencia como atributos relevantes de la señal de corriente, incluyendo el error cuadrático medio (RMSE), medidas normalizadas y métricas relacionadas con calidad de forma de onda y esfuerzo de conmutación. En el caso de Q-learning, se presentan análisis univariados, visualizaciones conjuntas mediante mapas de calor, y un ranking de combinaciones que permite seleccionar controladores representativos (mejor, intermedio y peor). Para DQN, el estudio se desarrolla de manera progresiva en fases, combinando barridos amplios, ajustes de dinámica de entrenamiento y análisis univariados, además de una comparación de funciones de recompensa.

Los resultados de este capítulo permiten justificar la selección de agentes que posteriormente serán implementados y evaluados en *PLECS Standalone*, estableciendo así el punto de partida para los análisis de desempeño bajo condiciones de operación más cercanas a la aplicación final.

5.1. Métricas de desempeño

A continuación, se definen las métricas utilizadas para evaluar el desempeño de los controladores implementados. Cada una permite cuantificar distintos aspectos del seguimiento de la señal de referencia y la calidad de la forma de onda resultante.

- **Error cuadrático medio (RMSE):** Representa el valor cuadrático medio del error y para mediciones discretas se puede calcular como:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(i_{\mathrm{ref}}(k) - i(k))^2} \tag{5.1}$$

  donde N es el número total de muestras, $i_{\mathrm{ref}}(k)$ es la corriente de referencia e $i(k)$ la corriente medidas en el instante k.

- **Error cuadrático medio relativo (RMSE %):** Representa el valor cuadrático medio del error, normalizado respecto a la amplitud de la referencia y expresado en porcentaje.

$$\mathrm{RMSE}\,\% = \frac{\mathrm{RMSE}}{A_{\mathrm{ref}}} \times 100\,\% \tag{5.2}$$

  donde A_{ref} la amplitud de la corriente de referencia.

- **Distorsión armónica total (THD):** Cuantifica la proporción de contenido armónico respecto al componente fundamental de la corriente medida, expresada en porcentaje:

$$\mathrm{THD}\,\% = \frac{\sqrt{\sum_{n=2}^{\infty} I_n^2}}{I_1} \times 100\,\% \tag{5.3}$$

  donde I_n es el valor RMS del n-ésimo armónico y I_1 el valor RMS del componente fundamental.

- **Ganancia relativa de la componente fundamental:** Evalúa la relación entre la amplitud del componente fundamental de la corriente medida y la amplitud de la referencia, expresada en porcentaje.

$$G[\,\%] = \frac{A_1}{A_{\mathrm{ref}}} \times 100\,\% \tag{5.4}$$

  donde A_1 corresponde a la amplitud de la componente fundamental de la corriente medida.

- **Frecuencia de conmutación (f_{sw}):** Representa la frecuencia promedio, expresada en kHz, con la que el inversor cambia el estado de sus interruptores durante un intervalo de tiempo determinado:

$$f_{\mathrm{sw}} = \frac{n_1 + n_2}{2T_{\mathrm{ciclo}}} \tag{5.5}$$

  donde n_1 y n_2 corresponden al número de conmutaciones registradas de cada interruptor durante un ciclo T_{ciclo}.

5.2. Control de corriente monofásico con agente Q-learning

En esta sección se presentan los resultados obtenidos a partir del estudio de los parámetros del algoritmo Q-learning aplicado al control de corriente monofásico detallado en la sección 3.3 del Capítulo 3. El esquema de control y entrenamiento del agente utilizado se muestra en la Figura 3.2 del capítulo antes mencionado.

Cada configuración de parámetros fue evaluada utilizando una misma secuencia de señales de referencia y el desempeño del agente se cuantificó mediante el error cuadrático medio (Ecuación 5.1). De este modo, la selección del conjunto óptimo de hiperparámetros se basa exclusivamente en la calidad del seguimiento de corriente bajo condiciones dinámicas idénticas para todos los agentes.

Para todas las simulaciones se utilizó el mecanismo de *Action Hold* (AH) descrito en el Apéndice A. Este mecanismo, que mantiene la acción durante un número fijo de pasos de simulación, resultó necesario para evitar cambios excesivamente rápidos en la conmutación y para permitir que el agente estabilizara la política frente a señales con transiciones bruscas. Salvo que se indique lo contrario, los resultados presentados en esta sección consideran un valor de AH = 5, que entregó el desempeño más estable durante las pruebas preliminares.

A continuación se muestran los análisis realizados sobre los 441 agentes entrenados, incluyendo un estudio univariado para cada hiperparámetro, mapas de calor del desempeño, la identificación de los mejores y peores controladores obtenidos, así como una comparación cualitativa entre casos representativos. Finalmente, se presenta el conjunto de parámetros seleccionado como óptimo y se discuten las principales observaciones derivadas de estos resultados.

5.2.1. Análisis univariado de los hiperparámetros

En este apartado se analiza de forma independiente la influencia de cada hiperparámetro del algoritmo Q-learning sobre el RMSE alcanzado. Para ello, se agruparon los agentes según el valor del parámetro en estudio y se calculó la mediana del RMSE, siguiendo lo descrito en el Apéndice D, dado que en ciertas configuraciones de hiperparámetros el control presenta divergencias. Por esto, el uso del RMSE promedio puede verse fuertemente afectado por valores atípicos, induciendo interpretaciones sesgadas de los resultados. Por este motivo, la mediana constituye una medida más representativa del comportamiento típico del agente bajo variabilidad estocástica [25].

5.2.1.a. Efecto de la tasa de aprendizaje

La tasa de aprendizaje α muestra una influencia directa en el desempeño del agente Q-learning. Como se observa en la Figura 5.1, los valores bajos de α, particularmente $\alpha = 0{,}1$, producen un RMSE considerablemente mayor en comparación con el resto de los valores evaluados, lo que sugiere un aprendizaje más lento y menos estable.

A partir de $\alpha = 0{,}2$ se aprecia una disminución significativa del RMSE en comparación con $\alpha = 0{,}1$, y para valores moderados y altos ($0{,}5 \leq \alpha \leq 0{,}9$) el desempeño se mantiene dentro de un rango estrecho, indicando una mayor consistencia en el proceso de aprendizaje. En este entorno, tasas de aprendizaje elevadas permiten al agente actualizar los valores Q con suficiente rapidez para capturar la dinámica del sistema sin inducir inestabilidades en la política aprendida.

En consecuencia, $\alpha = 0{,}1$ debe descartarse debido a su alto valor de mediana del RMSE, mientras que todos los valores desde $\alpha = 0{,}4$ en adelante muestran un desempeño adecuado y estable para el problema de control considerado.

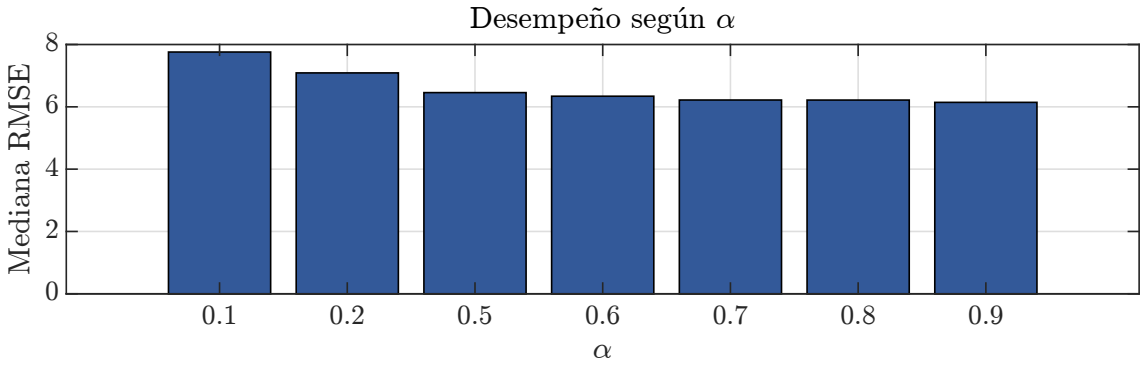


Figura 5.1: Mediana RMSE para cada valor de la tasa de aprendizaje α.

5.2.1.b. Efecto del factor de descuento

El factor de descuento γ presenta un efecto notable sobre el desempeño del agente Q-learning. Tal como se aprecia en la Figura 5.2, los valores más bajos de γ, producen los menores valores de RMSE promedio, con diferencias muy reducidas entre ellos, lo que indica un comportamiento estable y un aprendizaje adecuado para este entorno.

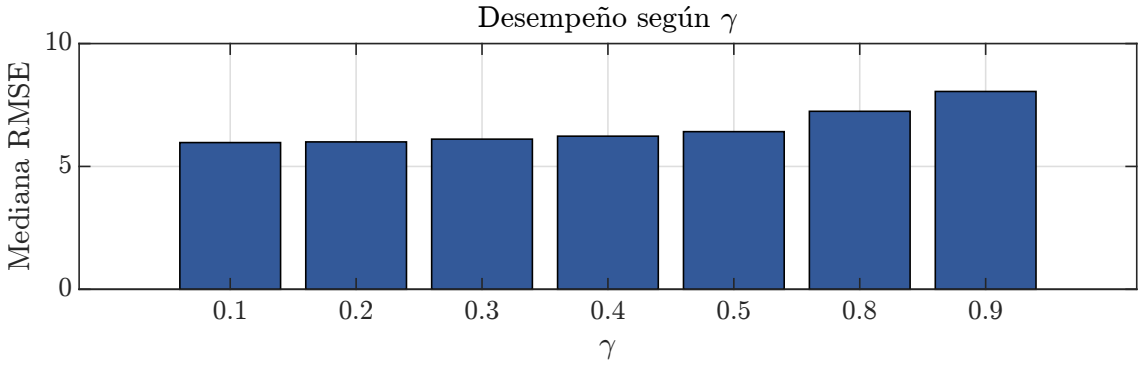


Figura 5.2: Mediana RMSE para cada valor del factor de descuento γ.

Por el contrario, valores altos del factor de descuento ($\gamma = 0{,}8$ y $\gamma = 0{,}9$) incrementan el RMSE promedio de manera significativa, lo que evidencia un deterioro del desempeño. Esto indica que el agente pierde precisión cuando otorga demasiado peso al largo plazo, posiblemente debido a la interacción entre la discretización del estado, la naturaleza local de la recompensa y la necesidad de reacciones rápidas ante cambios en la referencia.

En conjunto, los resultados muestran que los valores de $\gamma \in [0{,}1,\ 0{,}5]$ son los más adecuados para el problema estudiado, mientras que valores superiores deben descartarse debido al aumento consistente del error de seguimiento.

5.2.1.c. Efecto de la exploración

La probabilidad de exploración ε muestra una influencia significativamente menor en el desempeño del agente en comparación con los otros hiperparámetros evaluados. Como se aprecia en la Figura 5.3, los tres valores considerados ($\varepsilon = 0{,}01$, $0{,}05$ y $0{,}1$) producen RMSE promedio muy similares, sin diferencias apreciables que indiquen un efecto destacado sobre la calidad del segui-

miento de corriente. Este comportamiento sugiere que, para el rango de parámetros explorado y bajo la función de recompensa utilizada, el agente logra un equilibrio adecuado entre exploración y explotación aun con niveles de exploración relativamente bajos. El entorno monofásico estudiado presenta transiciones rápidas y una señal de recompensa altamente local, lo que reduce la necesidad de una exploración amplia para descubrir políticas efectivas. En consecuencia, ninguno de los valores evaluados para ε introduce un deterioro o mejora significativa en el desempeño, por lo que este hiperparámetro puede considerarse menos crítico en comparación con α y γ.

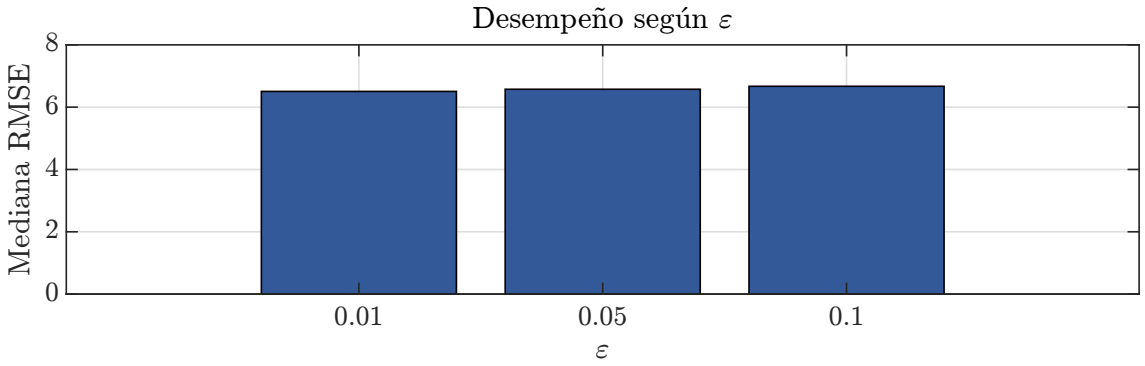


Figura 5.3: RMSE promedio para cada valor de la probabilidad de exploración ε.

5.2.1.d. Efecto del paso de discretización

El paso de discretización del error e_{step} tiene un efecto perceptible en el desempeño del agente Q-learning. Como se observa en la Figura 5.4, un paso pequeño ($e_{\text{step}} = 0{,}1$), que en este caso implica un espacio de estados de 801 niveles, conduce al mayor RMSE promedio entre los tres valores evaluados. Esto indica que una resolución demasiado fina dificulta el aprendizaje, pues incrementa de forma significativa el número de estados que el agente debe explorar y actualizar.

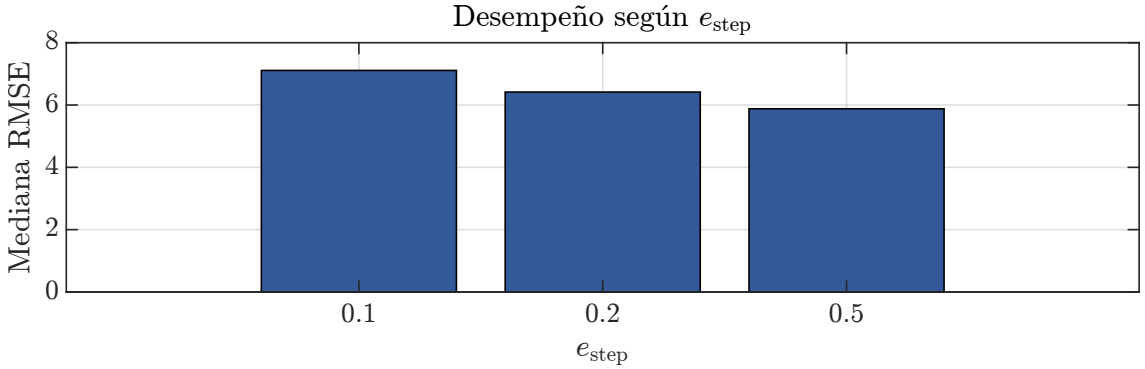


Figura 5.4: Mediana RMSE para cada valor del paso de discretización e_{step}.

Por el contrario, los valores intermedios y gruesos de discretización ($e_{\text{step}} = 0{,}2$ y $0{,}5$, correspondientes a 401 y 161 niveles del espacio de estados, respectivamente) reducen el valor de la medianda y promedio del RMSE mostrando un comportamiento más estable. Estos tamaños más compactos permiten agrupar estados similares y favorecen la generalización, simplificando la tabla Q y facilitando la convergencia del aprendizaje. Para este entorno monofásico, donde la dinámica es rápida y suave, esta reducción en la granularidad no compromete la precisión del control, sino que la mejora.

Por lo tanto, se puede decir que los resultados indican que los valores de e_{step} entre 0.2 y 0.5 proporcionan un equilibrio más adecuado entre la complejidad del espacio de estados y la capacidad de aprendizaje del agente, mientras que $e_{\text{step}} = 0{,}1$ debe evitarse debido a su mayor valor de mediana del RMSE.

5.2.1.e. Conclusiones

A partir del análisis univariado de los hiperparámetros del algoritmo Q–learning, es posible extraer diversas conclusiones relevantes sobre la sensibilidad del método y su comportamiento bajo distintas configuraciones. Se presenta la Tabla 5.1, que contiene los valores estadísticos correspondientes a todas las combinaciones de hiperparámetros exploradas.

Valor	**N° datos**	**RMSE**			
		Promedio	Mediana	Mínimo	Máximo
Variación de α					
0.1	63	7.9696	7.7587	6.4013	11.373
0.2	63	7.2411	7.0887	5.8428	9.8917
0.5	63	6.5827	6.4564	5.5545	8.8773
0.6	63	6.5108	6.3408	5.5095	8.7040
0.7	63	6.4535	6.2187	5.4301	8.6316
0.8	63	6.4381	6.2170	5.4540	8.5720
0.9	63	6.4550	6.1434	5.3195	8.5430
Variación de γ					
0.1	63	6.1587	5.9712	5.3195	7.9782
0.2	63	6.2593	5.9985	5.4793	8.1598
0.3	63	6.3372	6.1106	5.4606	8.2448
0.4	63	6.4624	6.2318	5.5256	8.5567
0.5	63	6.6203	6.4158	5.6223	8.8349
0.8	63	7.5417	7.2440	6.3102	10.430
0.9	63	8.2712	8.0509	6.9061	11.373
Variación de ϵ					
0.01	147	6.7414	6.5051	5.3195	11.373
0.05	147	6.7950	6.5743	5.4422	10.891
0.10	147	6.8853	6.6700	5.5856	11.030
Variación de e_{step}					
0.1	147	7.4600	7.1100	5.9712	11.373
0.2	147	6.7014	6.4158	5.5376	9.8389
0.5	147	6.2604	5.8799	5.3195	8.4048

Tabla 5.1: Resumen de estadísticas de RMSE para cada hiperparámetro del algoritmo Q-learning.

En síntesis, el análisis univariado revela que el desempeño del agente Q–learning es altamente sensible a determinados hiperparámetros, mientras que otros presentan una influencia secundaria. La tasa de aprendizaje α muestra un efecto particularmente marcado: valores elevados permiten una convergencia más rápida y estable, mientras que valores muy bajos derivan en un RMSE elevado. De manera similar, el factor de descuento γ presenta un comportamiento claro, donde valores bajos priorizan recompensas inmediatas y producen los menores errores, mientras que valores altos deterioran significativamente la calidad del seguimiento.

En contraste, la probabilidad de exploración ε exhibe un impacto reducido dentro del rango evaluado, lo que sugiere que el entorno monofásico y la estructura de la recompensa permiten un desempeño consistente incluso con niveles bajos de exploración. Finalmente, el paso de discretización e_{step} demuestra ser un hiperparámetro discretizaciones demasiado finas dificultan

el aprendizaje, mientras que valores intermedios o gruesos simplifican la tabla Q y favorecen la estabilidad del controlador.

5.2.2. Mapa de calor parámetros

El mapa de calor de la Figura 5.5 permite visualizar la interacción conjunta entre los hiperparámetros α y γ, evidenciando regiones de mejor y peor desempeño del agente Q-learning. En términos generales, el RMSE promedio disminuye de manera consistente al aumentar la tasa de aprendizaje, observándose los valores más bajos en el rango superior de α para casi todos los valores de γ. Esto confirma la tendencia previamente identificada en el análisis univariado, donde tasas de aprendizaje más altas favorecen una actualización eficiente de la tabla Q.

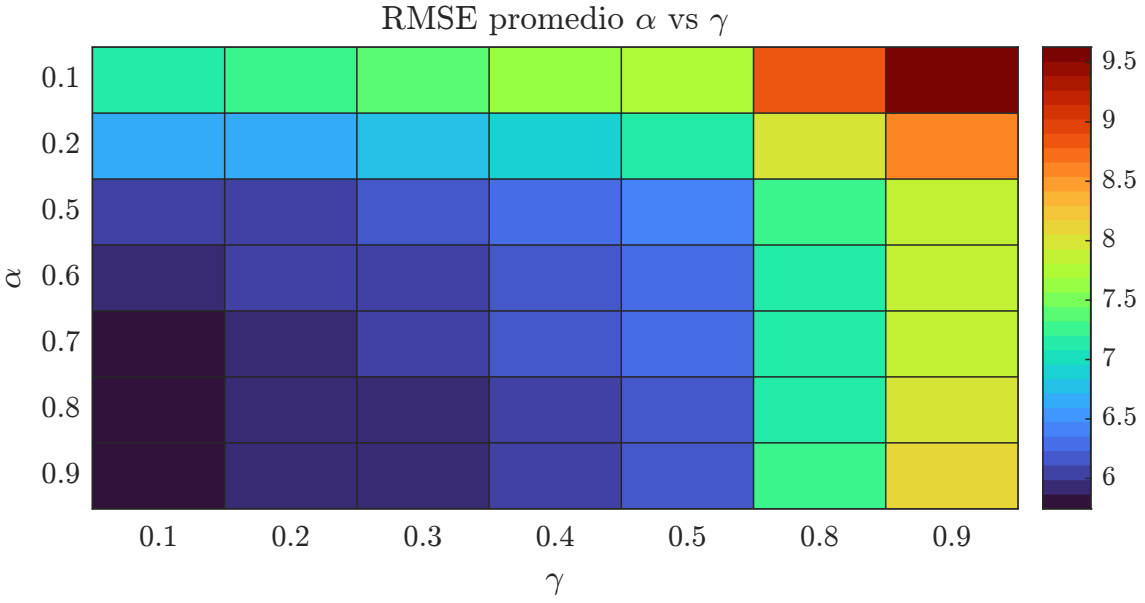


Figura 5.5: RMSE promedio del seguimiento de corriente en función de α y γ.

Por otro lado, el efecto del factor de descuento para cualquier valor de α, los menores valores de γ producen de manera sistemática los errores más bajos, mientras que los valores altos de descuento incrementan significativamente el RMSE. La degradación del desempeño al utilizar $\gamma = 0{,}8$ o $\gamma = 0{,}9$ es particularmente marcada en la parte superior del mapa, indicando que un horizonte temporal demasiado amplio perjudica la capacidad del agente para responder a la dinámica rápida del sistema monofásico.

La región óptima se localiza en la zona inferior izquierda del mapa, correspondiente a tasas de aprendizaje altas ($\alpha \geq 0{,}6$) combinadas con factores de descuento bajos ($\gamma \leq 0{,}3$). Esta combinación promueve un aprendizaje rápido y eficiente, centrado en recompensas inmediatas, lo que resulta adecuado para el control de corriente en sistemas donde la información relevante se renueva a gran velocidad.

5.2.3. Ranking de combinaciones evaluadas

La Tabla 5.2 muestra con claridad cuáles combinaciones de parámetros permiten obtener el mejor desempeño en términos de RMSE. Los agentes con menor error comparten una tendencia evidente:

utilizan tasas de aprendizaje altas ($\alpha \geq 0{,}6$), factores de descuento bajos o moderados ($\gamma \leq 0{,}4$) y una discretización más gruesa del error ($e_{\text{step}} = 0{,}5$). Esta configuración coincide con los patrones observados previamente en los análisis por parámetro y en el mapa de calor, lo que refuerza la idea de que el agente aprende de forma más estable cuando el espacio de estados es compacto y el énfasis está puesto en las recompensas inmediatas.

Tabla 5.2: Los 10 mejores y 10 peores resultados del entrenamiento Q-learning según RMSE.

Mejores						**Peores**					
Run	α	γ	ε	e_{step}	RMSE	Run	α	γ	ε	e_{step}	RMSE
381	0.9	0.1	0.01	0.5	5.32	55	0.1	0.9	0.01	0.1	11.37
255	0.7	0.1	0.01	0.5	5.43	61	0.1	0.9	0.1	0.1	11.03
384	0.9	0.1	0.05	0.5	5.44	58	0.1	0.9	0.05	0.1	10.89
318	0.8	0.1	0.01	0.5	5.45	46	0.1	0.8	0.01	0.1	10.43
399	0.9	0.3	0.01	0.5	5.46	52	0.1	0.8	0.1	0.1	10.27
327	0.8	0.2	0.01	0.5	5.48	49	0.1	0.8	0.05	0.1	10.09
390	0.9	0.2	0.01	0.5	5.50	118	0.2	0.9	0.01	0.1	9.89
210	0.6	0.3	0.01	0.5	5.51	121	0.2	0.9	0.05	0.1	9.88
408	0.9	0.4	0.01	0.5	5.53	124	0.2	0.9	0.1	0.1	9.86
330	0.8	0.2	0.05	0.5	5.53	62	0.1	0.9	0.1	0.2	9.84

Por el contrario, los agentes con peor rendimiento se agrupan en torno a valores bajos de α y altos de γ, junto con discretizaciones finas del error. Estas combinaciones, que también aparecían como problemáticas en los análisis anteriores, producen errores significativamente mayores y una convergencia más irregular, evidenciando que el agente tiene dificultades para explorar y actualizar adecuadamente una tabla Q demasiado grande.

Se seleccionaron tres agentes representativos para ilustrar diferencias cualitativas en el comportamiento dinámico: el mejor agente del conjunto, correspondiente a la combinación ($\alpha = 0{,}9,\ \gamma = 0{,}1,\ \varepsilon = 0{,}01,\ e_{\text{step}} = 0{,}5$) asociada al Run 381; un caso intermedio, dado por el agente del Run 346 con parámetros ($\alpha = 0{,}8,\ \gamma = 0{,}4,\ \varepsilon = 0{,}05,\ e_{\text{step}} = 0{,}1$); y finalmente el peor desempeño, representado por el Run 55 con la combinación ($\alpha = 0{,}1,\ \gamma = 0{,}9,\ \varepsilon = 0{,}01,\ e_{\text{step}} = 0{,}1$).

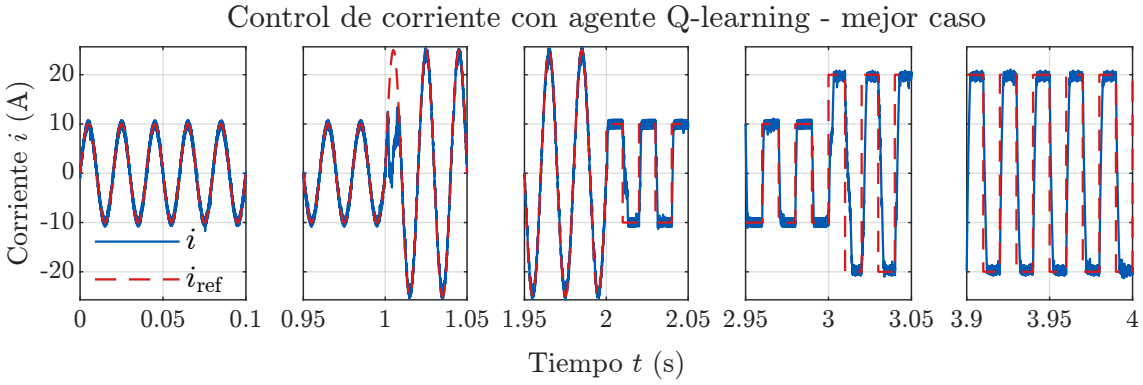


Figura 5.6: Control de corriente con el agente Q-learning de mejor desempeño (Run 381).

Los resultados del entrenamiento online muestran una marcada dependencia del desempeño

con respecto a los parámetros de aprendizaje. En el mejor caso, mostrado en la Figura 5.6, el agente logra un seguimiento estable y preciso de la referencia. La combinación de una tasa de aprendizaje alta y un factor de descuento bajo favorece actualizaciones rápidas y centradas en el error inmediato, mientras que la baja exploración contribuye a una política más consistente una vez estabilizado el aprendizaje. Cabe destacar que, al evaluar el mismo agente con AH = 1, el seguimiento de las señales cuadradas se volvió inestable (RMSE = 7.99), lo que motivó la adopción de AH = 5 para las simulaciones finales.

En el caso intermedio, ilustrado en la Figura 5.7, el agente mantiene un seguimiento razonable, pero con desviaciones más evidentes en las transiciones. Aunque utiliza una tasa de aprendizaje elevada, la combinación con un factor de descuento mayor y una discretización más fina del error aumenta la complejidad del espacio de estados, ralentizando la convergencia y deteriorando parcialmente la precisión del seguimiento.

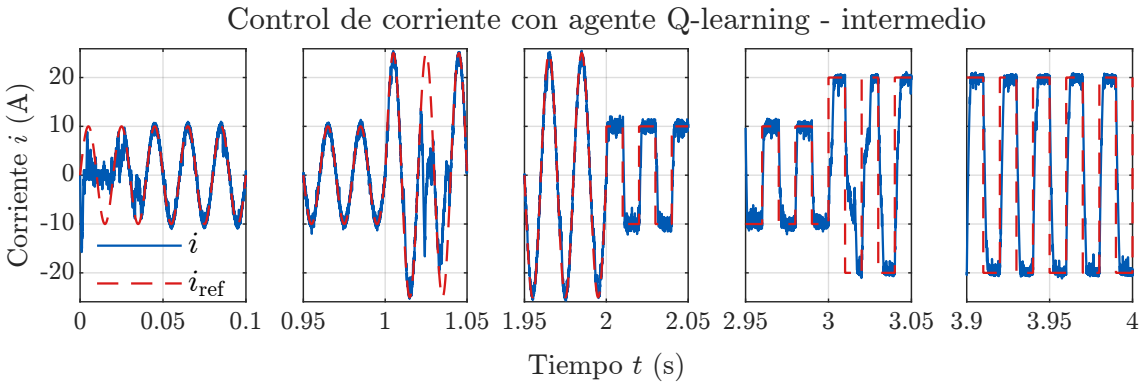


Figura 5.7: Control de corriente con un agente de desempeño intermedio (Run 346).

El peor caso, mostrado en la Figura 5.8, evidencia la dificultad del agente para aprender una política efectiva bajo una tasa de aprendizaje baja y un factor de descuento excesivamente alto. En esta configuración, el agente da prioridad a recompensas futuras poco relevantes para la dinámica rápida del sistema monofásico, lo que retrasa significativamente la adquisición de una política estable. Esto se refleja en transitorios amplios, incapacidad de seguir adecuadamente la referencia y, en algunos intervalos, una caída prematura hacia acciones poco útiles antes de que el aprendizaje converge.

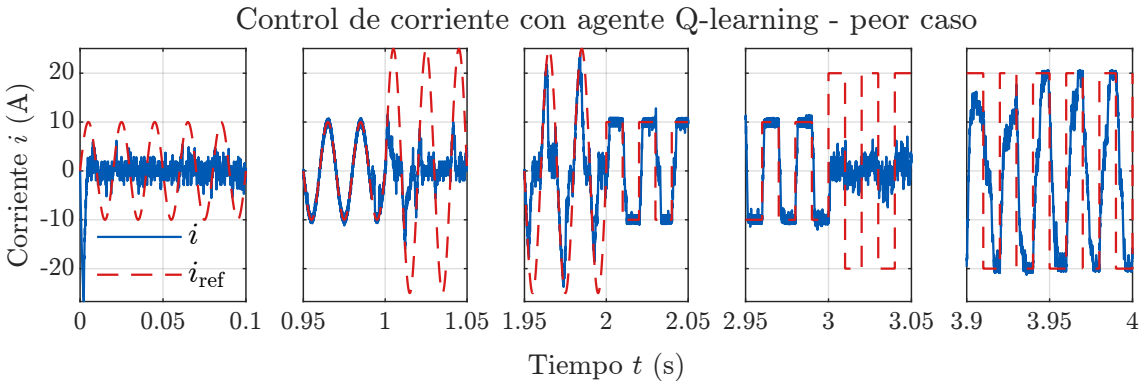


Figura 5.8: Control de corriente con el agente Q-learning de peor desempeño (Run 55).

5.3. Control de corriente monofásico con agente DQN

El entrenamiento del agente DQN, siguiendo el esquema mostrado en la Figura 3.3, se realizó utilizando episodios breves diseñados para capturar el comportamiento fundamental de seguimiento de corriente sin incurrir en tiempos de simulación excesivos. Cada episodio corresponde a dos ciclos de una señal senoidal de $50\,\text{Hz}$, con una duración total de $40\,\text{ms}$. El número exacto de pasos por episodio depende del tiempo de muestreo adoptado en cada fase del estudio; en la primera instancia, se utilizó un paso de $T_s = 10^{-5}$ s, lo que determina la cantidad de interacciones agente–entorno durante cada ejecución.

Para favorecer la capacidad de generalización del agente, la amplitud de la referencia se varió entre episodios dentro de un rango amplio, de modo que el controlador aprendiera una política que no dependiera de un único nivel de operación. La recompensa empleada inicialmente se definió como

$$r = 1 - e_N^2, \tag{5.6}$$

donde e_N corresponde al error normalizado. Esta formulación entrega una escala unificada para comparar el efecto de nuevos términos o ajustes introducidos en etapas posteriores del entrenamiento.

El fin de cada episodio se determinó únicamente por el número máximo de pasos que completan los $40\,\text{ms}$ especificados, de modo que no se utilizó criterio de terminación adicional.

Siguiendo la metodología mencionada en el Capítulo 3, se presentan los resultados obtenidos de manera progresiva para encontrar un conjunto de hiperparámetros y configuración para realizar un control de corriente adecuado.

5.3.1. Fase I: Exploración Amplia del espacio de Hiperparámetros

Con el propósito de obtener una primera caracterización del comportamiento del agente DQN, se realizó un barrido inicial sobre los hiperparámetros listados en la Tabla 5.3. Esta exploración permitió identificar configuraciones capaces de producir un aprendizaje estable bajo un tiempo de muestreo de $T_s = 10^{-5}$ s y episodios de duración fija $T_{\text{ep}} = 40\,\text{ms}$ (equivalentes a dos ciclos de una señal senoidal de $50\,\text{Hz}$), tal como se describió en la introducción del entrenamiento.

Tabla 5.3: Parámetros considerados en la exploración del primer barrido DQN.

Parámetro	Valores
Neuronas por capa N_n	24, 32
Tasa de aprendizaje α	5e-4, 1e-3, 5e-3
Factor de descuento γ	0.95, 0.97, 0.99
Decaimiento de exploración $\varepsilon_{\text{decay}}$	3.07e-5, 1.54e-5 ,1.02e-5
Target Smooth Factor (TSF)	1e-3
Tamaño de miniBatch $\lvert\mathcal{M}\rvert$	64, 128

El número total de combinaciones resultante de estos valores es$N_{\text{comb}} = 2\times3\times3\times3\times1\times2 = 108$.

Es importante destacar que el parámetro $\varepsilon_{\text{decay}}$ no se fijó como un valor único dentro del barrido: en su lugar, fue calculado para cada caso particular utilizando los episodios mínimos deseados y el número de pasos por episodio. En esta etapa se utilizaron tres configuraciones equivalentes a $E_\varepsilon = 80$, 160 y 242, que conducen respectivamente a los valores de decaimiento mostrados

en la Tabla 5.3. Esto garantiza que la fase de exploración disminuya a ritmos compatibles con la duración efectiva del entrenamiento en cada configuración.

Finalizado el entrenamiento, todos los agentes fueron evaluados en un entorno de validación sin exploración, de manera que el desempeño reflejara exclusivamente la política aprendida. Para esta evaluación se utilizaron dos señales de referencia: una señal senoidal de 20 A y 50 Hz aplicada durante tres ciclos, y una señal cuadrada con las mismas características de amplitud y período. Dado que ambas comparten la misma frecuencia fundamental, sus errores pueden combinarse directamente en una métrica única basada en su contribución conjunta al error cuadrático medio.

La herramienta *Training Monitor* del *RL toolbox* permite observar la evolución de la recompensa promedio de los diferentes agentes durante el entrenamiento. En la Figura 5.9 se muestra la evolución de la recompensa promedio de los diferentes agentes entrenados con las combinaciones mostradas en la Tabla 5.3.

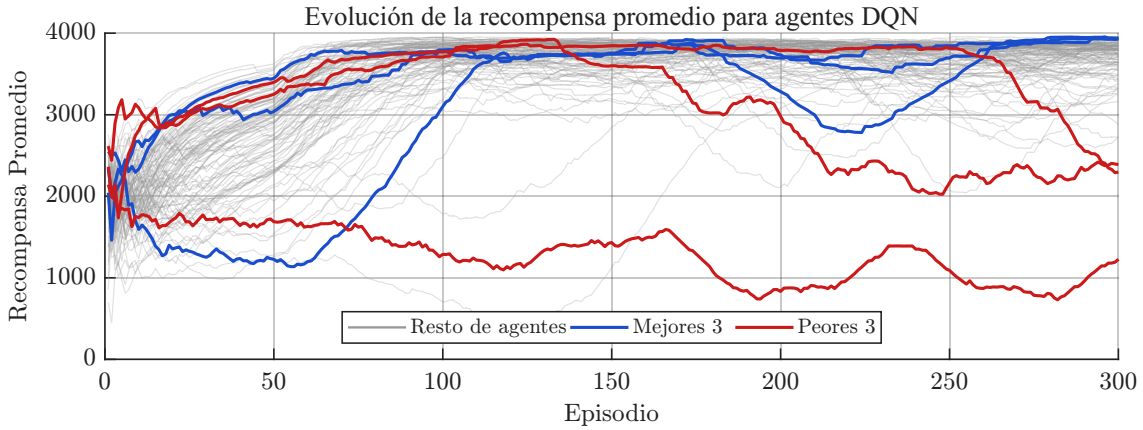


Figura 5.9: Recompensa promedio de entrenamientos de agentes en primer barrido de parámetros.

En la Figura 5.9 se presentan los agentes que obtuvieron las *mejores* y *peores* recompensas promedio al final del entrenamiento. Sin embargo, este resultado no permite concluir directamente si dichos agentes corresponden a los mejores o peores en términos del desempeño esperado del sistema, entendido como la minimización del RMSE de la corriente medida.

Esto se debe a que la recompensa es una métrica asociada al proceso de aprendizaje y refleja el progreso del entrenamiento del agente, mientras que el RMSE evalúa la calidad efectiva del seguimiento de corriente bajo una política fija. En consecuencia, ambos criterios responden a objetivos distintos y pueden inducir a ordenamientos diferentes entre los agentes evaluados.

No obstante, desde la perspectiva del entrenamiento, resulta deseable observar una baja variabilidad en la recompensa promedio entre distintos agentes. Esto sugiere un proceso de aprendizaje más estable y consistente, en el cual los agentes convergen hacia políticas de desempeño similar.

Tabla 5.4: Resultados del entrenamiento DQN según RMSE para primer barrido de parámetros.

Mejores							**Peores**						
Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	$\varepsilon_{\text{decay}}$	RMSE	Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	$\varepsilon_{\text{decay}}$	RMSE
078	32	5e-4	0.97	64	1.e-5	10.18	098	32	1e-4	0.99	64	1.5e-5	72.23
052	24	1e-4	0.99	128	3.1e-5	10.38	087	32	5e-4	0.97	128	1.e-5	63.87
099	32	1e-4	0.99	64	1.e-5	10.46	021	24	5e-4	0.95	64	1.e-5	63.06
002	24	1e-3	0.95	64	1.5e-5	10.56	091	32	1e-4	0.95	64	3.1e-5	61.78
016	24	1e-3	0.99	128	3.1e-5	10.61	033	24	5e-4	0.97	128	1.e-5	61.13

La Tabla 5.4 presenta los cinco mejores y cinco peores agentes del barrido DQN según el RMSE total obtenido en la etapa de evaluación. Los mejores resultados presentan valores de RMSE cercanos a 10 A, asociados principalmente a configuraciones con tamaños de minibatch moderados ($|\mathcal{M}| = 64$–128), tasas de aprendizaje bajas y factores de descuento elevados. En contraste, los peores agentes exhiben errores entre 60 y 72 A, aun cuando utilizan combinaciones de parámetros similares, lo que sugiere una alta sensibilidad del entrenamiento temprano a la estabilidad numérica y a la exploración. Estos resultados confirman que, en esta etapa inicial, el DQN muestra un comportamiento altamente dependiente de las condiciones de entrenamiento y del ajuste fino de los hiperparámetros.

Como ya se mencionó, es importante destacar que los agentes con mejor o peor desempeño según RMSE (Tabla 5.4) no coinciden con los agentes identificados como *mejores 3* y *peores 3* en el Figura 5.9 de recompensas del entrenamiento (Runs 40, 83 y 103 para las mejores curvas; Runs 37, 5 y 21 para las peores). Esta discrepancia se explica porque la recompensa refleja el proceso de aprendizaje durante los episodios, mientras que el RMSE evalúa la calidad del seguimiento de corriente bajo una política fija y sin exploración. En otras palabras, un agente puede mostrar una recompensa final alta sin que ello garantice un seguimiento preciso en la etapa de validación, y viceversa. Esta diferencia entre ambas métricas justifica el análisis basado en RMSE como criterio principal para seleccionar el agente final a homologar en *PLECS*.

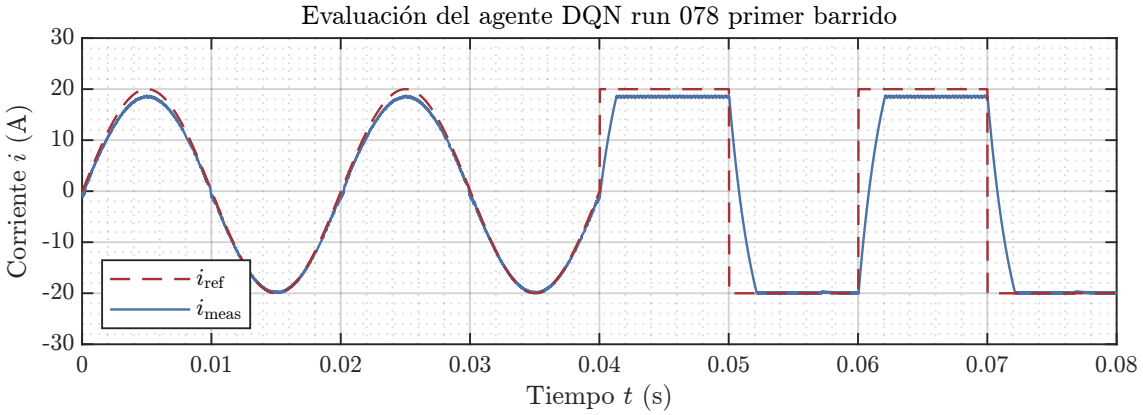


Figura 5.10: Evaluación de agente DQN con mejor desempeño (RMSE) en primer barrido.

La Figura 5.10 muestra la evaluación del agente DQN con mejor desempeño, correspondiente al primer barrido de hiperparámetros. En este caso, aunque el seguimiento de la referencia es adecuado gran parte del tiempo, pierde precisión en los valores máximos. Estos efectos son consistentes con un agente que todavía no alcanza una política suficientemente refinada, dado

que dicho agente no ha logrado generalizar el seguimiento para todos los estados.

5.3.2. Fase II: Ajuste de la dinámica del entorno y estabilidad de aprendizaje

Para evaluar el aprendizaje de los agentes durante un mayor número de episodios, se redujo la frecuencia de muestreo desde 100 kHz hasta 40 kHz, con el fin de disminuir la cantidad de pasos de entrenamiento sin perder la esencia de la dinámica del sistema. En la Figura 5.11 se muestra la evolución de la recompensa promedio obtenida por los agentes en este segundo barrido. Para este experimento se mantuvo un parámetro $\varepsilon_{\text{decay}}$ tal que la exploración decae hasta su valor mínimo en torno a los 300 episodios. Además, se incluyeron arquitecturas con 64 neuronas por capa para ampliar el espacio de búsqueda.

De manera general, se observa que la recompensa promedio de todos los agentes aumenta rápidamente hasta converger a valores similares antes de los 200 episodios, período en el cual predomina la fase de mayor exploración. Este comportamiento sugiere que una exploración más amplia permite almacenar muestras más representativas de la dinámica del sistema, lo que acelera el proceso de aprendizaje de la red neuronal.

Posterior a los 300 episodios, cuando la exploración alcanza su valor mínimo ($\varepsilon = 0{,}01$), varios agentes comienzan a perder la recompensa previamente lograda. Esto ocurre porque, si un agente no consiguió generalizar adecuadamente los distintos estados durante la fase de exploración, al pasar a la fase de explotación tiende a reforzar decisiones subóptimas y, en consecuencia, *olvida* el comportamiento que conducía a mayores recompensas.

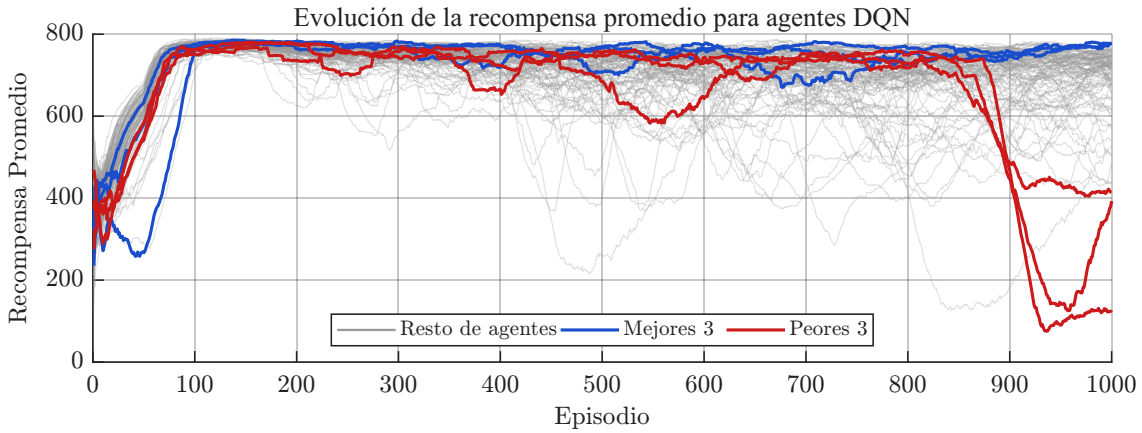


Figura 5.11: Recompensa promedio para entrenamiento de agentes DQN en 1000 episodios.

La Figura 5.12 corresponde al agente con mejor desempeño obtenido en el barrido de 1000 episodios. En comparación con el mejor caso del primer barrido de parámetros (Figura 5.10), se observa que el agente sigue la referencia con un pequeño *offset* a lo largo de toda la evaluación. Asimismo, se aprecia cierta irregularidad en la forma de onda de la corriente cuando la referencia senoidal cruza por cero. Adicionalmente, la disminución de la frecuencia de muestreo provoca un leve aumento visible en el rizado de la señal.

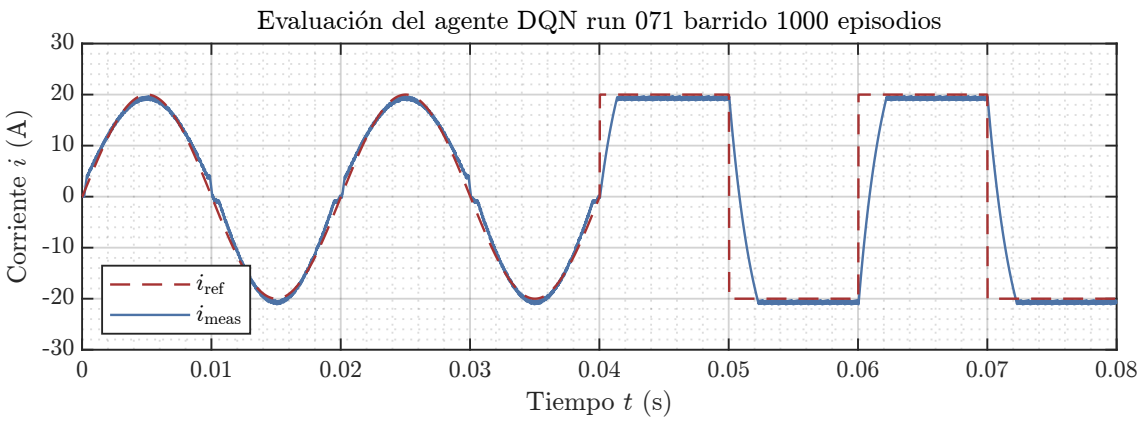


Figura 5.12: Evaluación de agente DQN con mejor desempeño (RMSE) en barrido de mil episodios.

Al comparar el RMSE total entre el mejor caso del agente del primer barrido (Tabla 5.4) y el mejor caso del agente de este barrido (Tabla B.1), se obtiene un incremento del 1.67 % en el caso del barrido de 1000 episodios. Esto es esperable, ya que, aunque la evaluación utiliza el mismo tiempo de muestreo, durante el entrenamiento el agente aprendió a tomar decisiones en intervalos temporales ligeramente mayores. Aun cuando el RMSE aumenta, la diferencia es acotada, lo que sugiere que, en etapas tempranas de exploración de hiperparámetros, es posible obtener una buena aproximación inicial incluso trabajando con frecuencias de muestreo algo menores.

5.3.3. Fase III: Análisis Univariado y Evaluación de Funciones de Recompensa

Tras las conclusiones obtenidas en las dos etapas previas, se diseñó una tercera fase orientada a estudiar de manera aislada el efecto de cada hiperparámetro sobre el desempeño del agente. El objetivo de esta etapa fue acotar de forma sistemática las combinaciones más prometedoras, reduciendo la variabilidad observada en los barridos anteriores y delimitando así una región de búsqueda más estable para fases posteriores de optimización.

Para ello, los resultados de todos los entrenamientos realizados, registrando como métrica central el error medio cuadrático RMSE, se efectuó un análisis univariado, examinando para cada hiperparámetro la distribución del desempeño de los agentes cuando dicho parámetro varía de manera individual, manteniendo los demás dentro de rangos razonables. Este procedimiento permitió identificar tendencias claras en parámetros como el *learning rate*, el factor de descuento γ, el tamaño del *minibatch* y el decaimiento de ε, delimitando rangos en los que el aprendizaje resulta más estable y eficiente.

Adicionalmente, se llevó a cabo un experimento orientado a comparar distintas funciones de recompensa con el fin de evaluar su impacto en la estabilidad del entrenamiento y la capacidad de seguimiento de corriente. Se analizaron formulaciones cuadráticas, penalizaciones basadas en bandas, términos exponenciales y funciones que incorporan el costo por conmutación. Este estudio permitió identificar qué estructuras de recompensa favorecen un comportamiento más consistente del agente y contribuyen a una mejor generalización.

5.3.3.a. Tasa de aprendizaje

La Figura 5.13, muestra la mediana del RMSE obtenida para los agentes entrenados con distintas tasas de aprendizaje α. Se observa que los valores más pequeños de α (entre 10^{-4} y 10^{-3}) producen

un desempeño similar y corresponden a las configuraciones más estables dentro del conjunto evaluado. En contraste, un valor mayor de $\alpha = 5 \times 10^{-3}$ incrementa de forma significativa la mediana del RMSE, indicando que tasas de aprendizaje altas dificultan la convergencia de la red y generan políticas de control menos precisas. Estos resultados sugieren que el DQN es sensible al *learning rate* y que valores moderados favorecen un entrenamiento más fiable.

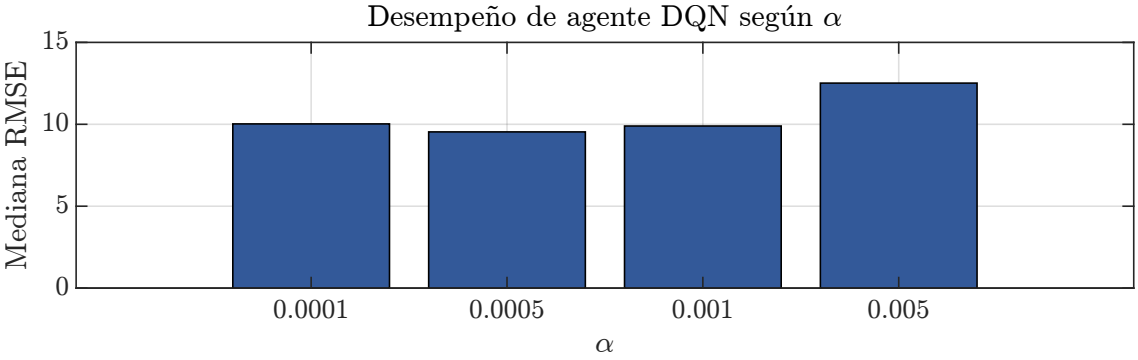


Figura 5.13: Desempeño de agentes DQN para tasa de aprendizaje α.

5.3.3.b. Factor de descuento

La Figura 5.14 muestra la mediana del RMSE obtenida para los agentes entrenados con distintos valores del factor de descuento γ. Se observa que los valores más bajos, en particular $\gamma = 0{,}96$, producen el mejor desempeño, mientras que factores de descuento más altos incrementan la mediana del error. Esto indica que, en este problema de control, otorgar demasiado peso a recompensas futuras puede dificultar la estabilización temprana de la política, generando un aprendizaje menos preciso. En consecuencia, valores moderados de γ resultan más adecuados para capturar correctamente la dinámica del sistema sin comprometer la estabilidad del entrenamiento.

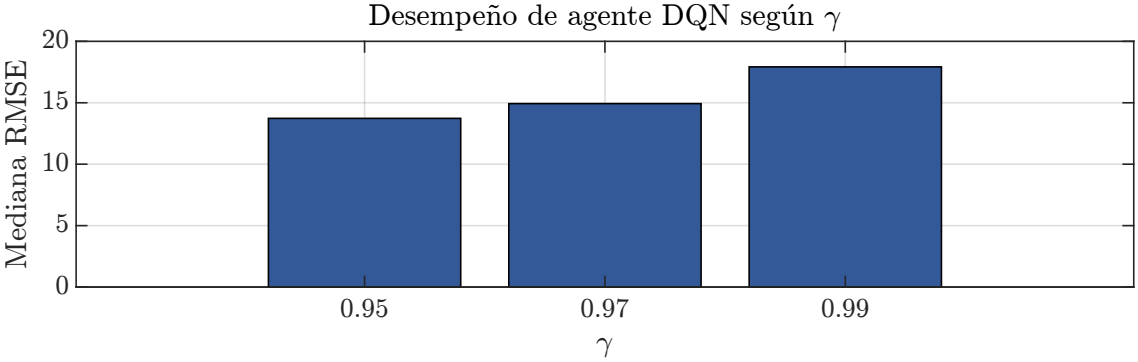


Figura 5.14: Desempeño de agentes DQN para el factor de descuento γ.

5.3.3.c. Número de neuronas por capa

La Figura 5.15, muestra la mediana del RMSE para agentes entrenados con diferentes números de neuronas por capa N_n. Se observa que la configuración intermedia de 32 neuronas ofrece el mejor desempeño, con una reducción ligera pero consistente del error respecto de las arquitecturas con 24 y 64 neuronas. Esto sugiere que aumentar el tamaño de la red no necesariamente mejora la calidad del control, ya que arquitecturas más grandes pueden dificultar la estabilización del

aprendizaje en esta etapa temprana, mientras que un tamaño moderado proporciona un equilibrio adecuado entre capacidad de representación y estabilidad del entrenamiento.

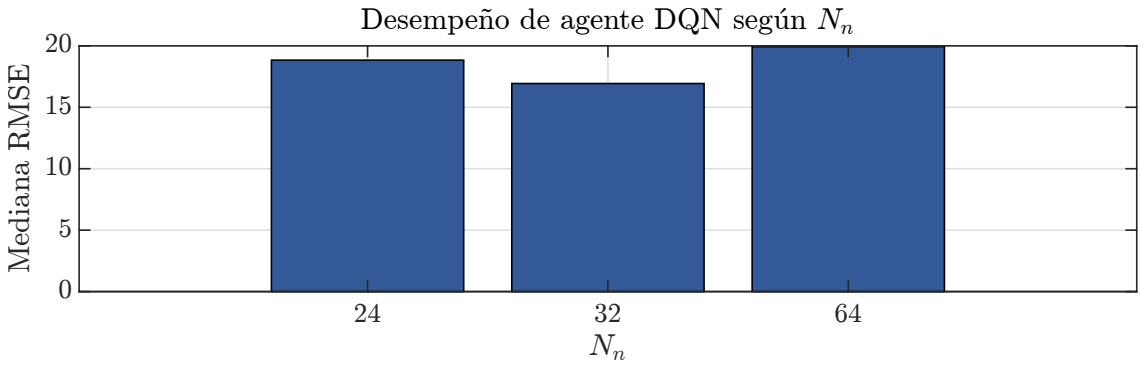


Figura 5.15: Desempeño de agentes DQN para distintos números de neuronas N_n por capa.

5.3.3.d. Tamaño del minibatch

La Figura 5.16 presenta la mediana del RMSE para distintos tamaños de minibatch $|\mathcal{M}|$. Se observa que los minibatch más pequeños (64 y 128) producen errores similares, mientras que el valor más grande ($|\mathcal{M}| = 256$) reduce de manera apreciable la mediana del RMSE. Esto indica que, en esta etapa del entrenamiento, un mayor tamaño de minibatch contribuye a estabilizar la estimación del gradiente y favorece la convergencia del agente. No obstante, el beneficio no es estrictamente lineal y su impacto debe considerarse en conjunto con otros parámetros, dado que minibatches excesivamente grandes pueden ralentizar la adaptación del agente frente a nuevas muestras del entorno.

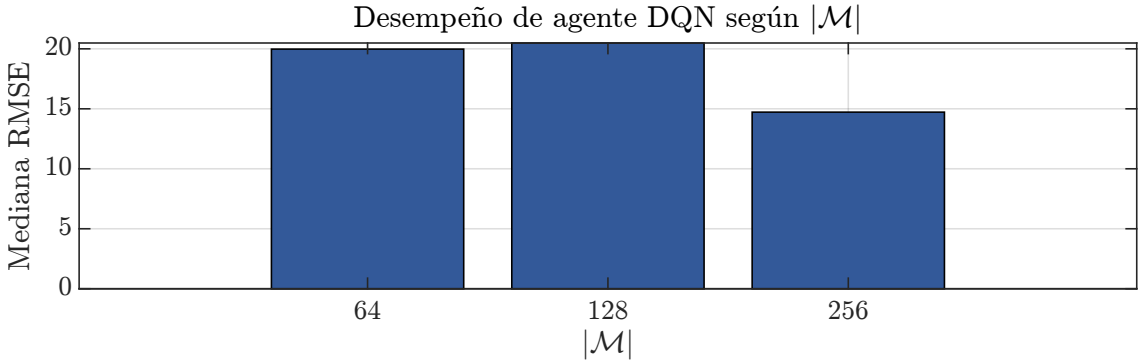


Figura 5.16: Desempeño de agentes DQN para el tamaño del minibatch $|\mathcal{M}|$.

5.3.3.e. Decaimiento del parámetro de exploración

La Figura 5.17 muestra la mediana del RMSE obtenida para distintos valores del decaimiento del parámetro de exploración $\varepsilon_{\text{decay}}$. Se observa que el menor valor evaluado ($1{,}02 \times 10^{-6}$) produce el mejor desempeño, mientras que valores más altos incrementan de forma consistente la mediana del error. Este resultado coincide con el hecho de que un $\varepsilon_{\text{decay}}$ más pequeño implica que la exploración se mantiene activa durante una mayor parte del entrenamiento, decayendo hasta su valor mínimo sólo hacia el final del proceso. En consecuencia, el agente dispone de más

oportunidades para recopilar experiencias diversas y representativas del entorno, lo que favorece una política final más estable y precisa.

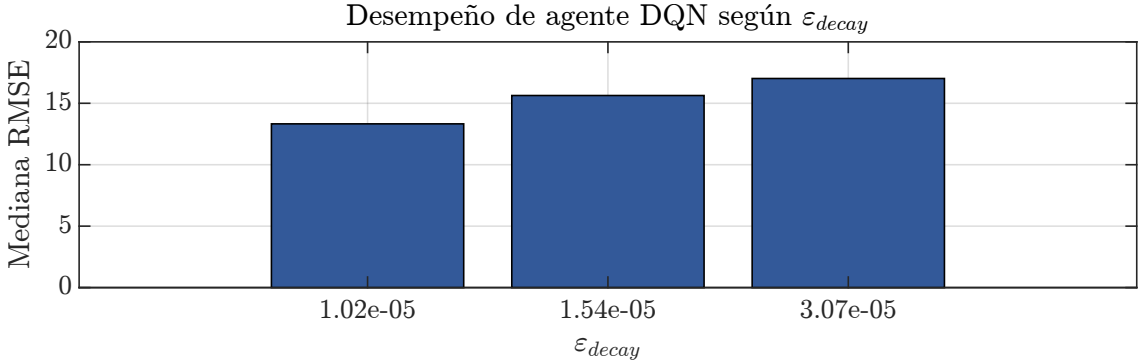


Figura 5.17: Desempeño de agentes DQN para el decaimiento del parámetro $\varepsilon_{\text{decay}}$.

5.3.3.f. Factor de suavizado del objetivo (*Target Smooth Factor*)

La Figura 5.18 muestra la mediana del RMSE obtenida para distintos valores del parámetro *Target Smooth Factor* (TSF). Se observa que el valor más pequeño (TSF $= 0{,}0001$) produce el peor desempeño, mientras que valores mayores disminuyen de manera significativa la mediana del error. Esto se debe a que un TSF muy bajo limita la actualización del valor objetivo en la red *target*, haciendo que la política evolucione de forma demasiado lenta y dificultando la adaptación del agente a las transiciones del entorno. En contraste, valores de TSF más altos permiten un suavizado menos restrictivo y favorecen una convergencia más rápida y estable, lo que se refleja en un RMSE menor.

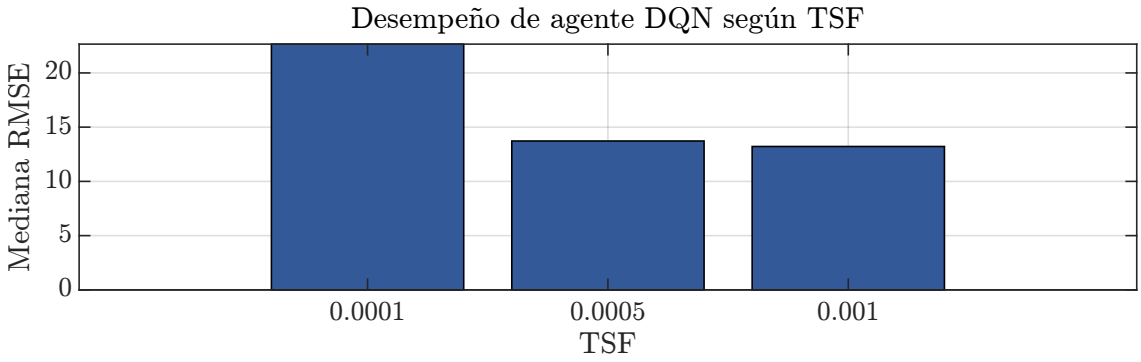


Figura 5.18: Desempeño de agentes DQN para el parámetro *Target Smooth Factor*.

5.3.3.g. Función de recompensa

La Figura 5.19 muestra la mediana del RMSE obtenida para los agentes entrenados con distintas funciones de recompensa. En este experimento se evaluaron cuatro formulaciones, todas basadas en el término principal $1 - e_N^2$, donde e_N corresponde al error de corriente normalizado. Las

funciones se definieron de la siguiente manera:

$$\text{Función 1:} \quad r_1 = 1 - e_N^2 \tag{5.7}$$

$$\text{Función 2:} \quad r_2 = r_1 + r_{\text{band}} \tag{5.8}$$

$$\text{Función 3:} \quad r_3 = r_1 + r_{\text{band}} + r_{\delta e} \tag{5.9}$$

$$\text{Función 4:} \quad r_4 = r_1 + r_{\text{band}} - \lambda|\Delta a| \tag{5.10}$$

Aquí, r_{band} corresponde a un incentivo por mantener el error dentro de una banda de control alrededor de la referencia; $r_{\delta e}$ premia la disminución del error respecto del paso anterior y castiga en caso contrario, favoreciendo transiciones más suaves; y $\lambda|\Delta a|$ introduce una penalización leve por conmutación, con el objetivo de desincentivar cambios innecesarios de acción.

Los resultados muestran que las funciones de recompensa que incorporan el término r_{band} (Funciones 2 y 4) tienden a producir las menores medianas de RMSE entre los agentes evaluados, lo que indica que incentivar explícitamente la permanencia del error dentro de una banda cercana a la referencia contribuye a un control más estable. La similitud en el desempeño de ambas funciones era esperable, dado que la penalización por conmutación incluida en la Función 4 es pequeña en comparación con el término principal del error y, por lo tanto, no modifica de manera significativa la señal de recompensa en esta etapa temprana del entrenamiento.

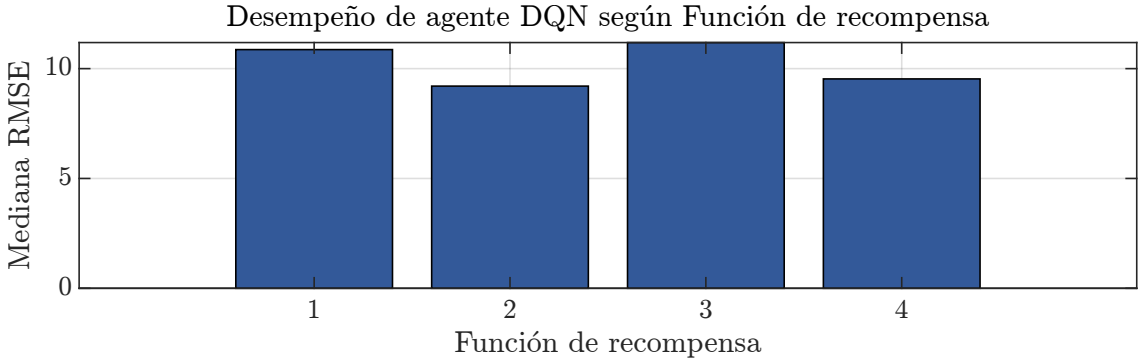


Figura 5.19: Desempeño de agentes DQN para distintas funciones de recompensa.

En contraste, la inclusión del término $r_{\delta e}$ en la Función 3 no mostró mejoras adicionales, lo que podría deberse a que premiar la reducción incremental del error introduce una señal más ruidosa y difícil de interpretar para el agente durante los primeros episodios de aprendizaje. Estos resultados sugieren que una función de recompensa simple, basada en el error cuadrático y un incentivo de banda, es suficiente para guiar al agente hacia un comportamiento estable durante los primeros barridos de hiperparámetros.

5.3.4. Entrenamiento Final

De manera transversal en los entrenamientos realizados se observa que los agentes tienden a obtener mejores resultados cuando la exploración se mantiene activa hasta etapas avanzadas del aprendizaje. Un decaimiento más lento de ε permite recopilar experiencias más diversas y representativas de la dinámica del sistema, lo que suele traducirse en políticas finales más estables.

Asimismo, el tamaño del *mini-batch* resulta relevante para la estabilidad del aprendizaje, ya que incorporar un conjunto mayor de transiciones en cada actualización reduce la varianza de la

estimación del gradiente y favorece una convergencia más consistente. Por otra parte, aumentar el tamaño de la red neuronal no garantiza un mejor desempeño: para el control de corriente monofásico considerado, arquitecturas compactas (por ejemplo, 24 o 32 neuronas por capa) suelen alcanzar resultados comparables o superiores, sugiriendo un equilibrio adecuado entre capacidad de representación y estabilidad del entrenamiento.

Cabe destacar que el entrenamiento presenta variabilidad inherente debido a la naturaleza estocástica del proceso, asociada tanto a la exploración como a la exposición del agente a referencias de distinta amplitud entre episodios. Por ello, más que esperar un único resultado determinista por configuración, se define un conjunto acotado de hiperparámetros base que permite comparar tendencias y seleccionar agentes según su desempeño en una etapa de evaluación sin exploración.

Tabla 5.5: Hiperparámetros de entrenamiento final.

Parámetro	**Valores**
Neuronas por capa N_n	24, 32
Tasa de aprendizaje α	1e-4, 5e-4, 1e-3, 5e-3
Factor de descuento γ	0.95, 0.97, 0.99

En todos los entrenamientos se mantuvieron constantes los siguientes parámetros: un factor de suavizado del objetivo TSF $= 5 \times 10^{-4}$, un tamaño de *mini-batch* de $|\mathcal{M}| = 256$, un decaimiento de la exploración $\varepsilon_{\text{decay}} = 1.54{\times}10^{-5}$ y una frecuencia de muestreo de $f_s = 50\,\text{kHz}$. Sobre esta configuración base se evaluaron los hiperparámetros variables del agente. La Tabla 5.5 resume el conjunto de combinaciones utilizadas en el entrenamiento final.

En la Figura 5.20, se muestra una evaluación del agente DQN con mejor desempeño según RMSE (Tabla 5.6). Los hiperparámetros elegidos son N_n = 32, α = 5e-4, γ = 0.99 con un RMSE = 8.90 A. Resultado similar obtuvo el agente con 24 neuronas por capa y manteniendo el resto de parámetros con un RMSE = 8.95 A durante los 80 ms de evaluación.

La evaluación del mejor agente obtenido en el barrido final muestra un seguimiento altamente preciso durante la etapa senoidal inicial la corriente medida se superpone casi exactamente con la referencia, sin desfases ni oscilaciones apreciables. En las transiciones hacia las referencias escalonadas el agente responde de manera rápida y estable, con errores transitorios mínimos debido a la respuesta natural del sistema resistivo inductivo, y en los niveles constantes mantiene un seguimiento firme.

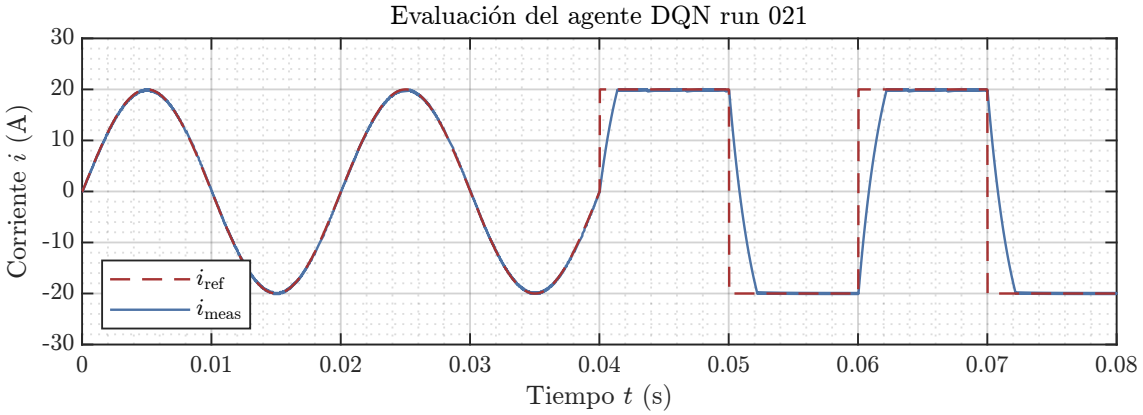


Figura 5.20: Evaluación de agente DQN con mejor desempeño (RMSE) en barrido final.

Es especialmente llamativo que, durante todo el entrenamiento, el agente haya sido expuesto exclusivamente a señales senoidales y, aun así hay agentes que logran un buen seguimiento frente a referencias cuadradas con cambios bruscos, lo que sugiere cierta capacidad de generalización aprendida de manera no explícita; sin embargo, este comportamiento no es universal, ya que en varios de los barridos de hiperparámetros se observaron agentes con un seguimiento deficiente ante este tipo de señales.

Tabla 5.6: Resultados entrenamiento final.

Mejores					**Peores**				
Run	N_n	α	γ	RMSE	Run	N_n	α	γ	RMSE
021	32	5.e-4	0.99	8.90	007	24	5.e-4	0.95	32.30
009	24	5.e-4	0.99	8.95	015	32	5.e-3	0.99	27.89
014	32	5.e-3	0.97	8.96	019	32	5.e-4	0.95	20.26
006	24	1.e-3	0.99	8.96	003	24	5.e-3	0.99	17.03
017	32	1.e-3	0.97	8.98	004	24	1.e-3	0.95	16.78

Capítulo 6

Análisis de controles de corrientes implementados en PLECS

En este capítulo se analiza el desempeño de los controladores basados en aprendizaje por refuerzo una vez implementados en la plataforma *PLECS Standalone*. A diferencia de los capítulos anteriores, donde el foco estuvo puesto en el entrenamiento y la selección de agentes, aquí se evalúa el comportamiento de las políticas finales bajo condiciones de operación representativas del sistema eléctrico.

Las pruebas consideradas incluyen el seguimiento de referencias de distinta forma y amplitud, variaciones en la frecuencia de la señal deseada y cambios en los parámetros de carga, permitiendo analizar tanto la precisión del control como su robustez frente a perturbaciones. Se presentan resultados para controladores monofásicos y trifásicos entrenados mediante Q-learning y *Deep Q-Network*, considerando distintas representaciones del sistema cuando corresponde.

El análisis desarrollado en este capítulo permite comparar de manera consistente el desempeño de las distintas estrategias de control en un entorno común de simulación, estableciendo las bases para discutir sus ventajas, limitaciones y posibles aplicaciones prácticas.

6.1. Control de corriente monofásico con agente Q-learning

A continuación se presentan diversas pruebas realizadas al control de corriente mediante un agente Q-learning. Los hiperparámetros seleccionados corresponden al mejor caso determinado en MATLAB: α = 0.9, γ = 0.1, ε = 0.01 y e_{step} = 0.5. En primera instancia, las pruebas se efectúan con el parámetro *Action Hold* AH = 1, lo que permite obtener un seguimiento más suave frente a referencias sinusoidales, para después evaluar el impacto de este parámetro y otros mecanismos estabilización de la política de acción mostrados en el Apendice A.

6.1.1. Control de corriente frente a cambio de resistencia

Se realizan cambios instantáneos en el valor de la resistencia de carga para analizar la respuesta del control frente a este tipo de perturbaciones. El experimento consiste en alcanzar primero un estado estacionario con una corriente de referencia de 10 A de amplitud y 50 Hz de frecuencia. Luego, en un instante determinado, se produce una disminución abrupta de la resistencia desde 5 Ω a 2 Ω (un 40 % de su valor inicial), y posteriormente se incrementa a 10 Ω (200 % del valor inicial). Como se observa en la Figura 6.1, durante todas las ventanas de tiempo se observa que la corriente medida mantiene el seguimiento de la referencia sin pérdida de estabilidad tras los cambios de R, con variaciones transitorias acotadas.

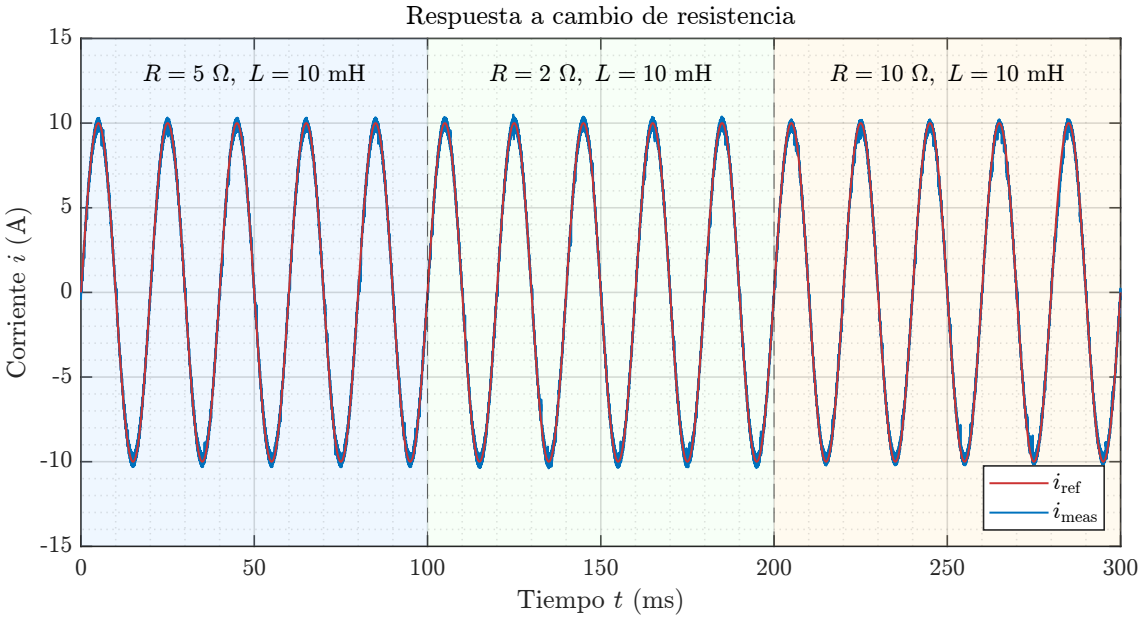


Figura 6.1: Respuesta del control Q-learning ante variaciones de la resistencia de carga en PLECS.

6.1.2. Control de corriente frente a cambio de inductancia

Al igual que en el experimento con cambio de resistencia, se realizan variaciones instantáneas en el valor de la inductancia para analizar la respuesta del control frente a este tipo de perturbaciones. Luego de alcanzar el estado estacionario con una corriente de referencia de 10 A de amplitud y 50 Hz de frecuencia, se produce una disminución abrupta de la inductancia desde 10 mH a 5 mH (un 50 % de su valor inicial) y, posteriormente, se incrementa a 15 mH (150 % del valor inicial).

Como se observa en la Figura 6.2, tras el primer cambio de inductancia el controlador presenta

dificultades para adaptarse a la nueva dinámica del sistema durante los primeros 50 ms, para luego lograr seguir la señal de referencia, aunque con un mayor rizado. Finalmente, en los 200 ms, al aumentar la inductancia a 15 mH, la señal se estabiliza.

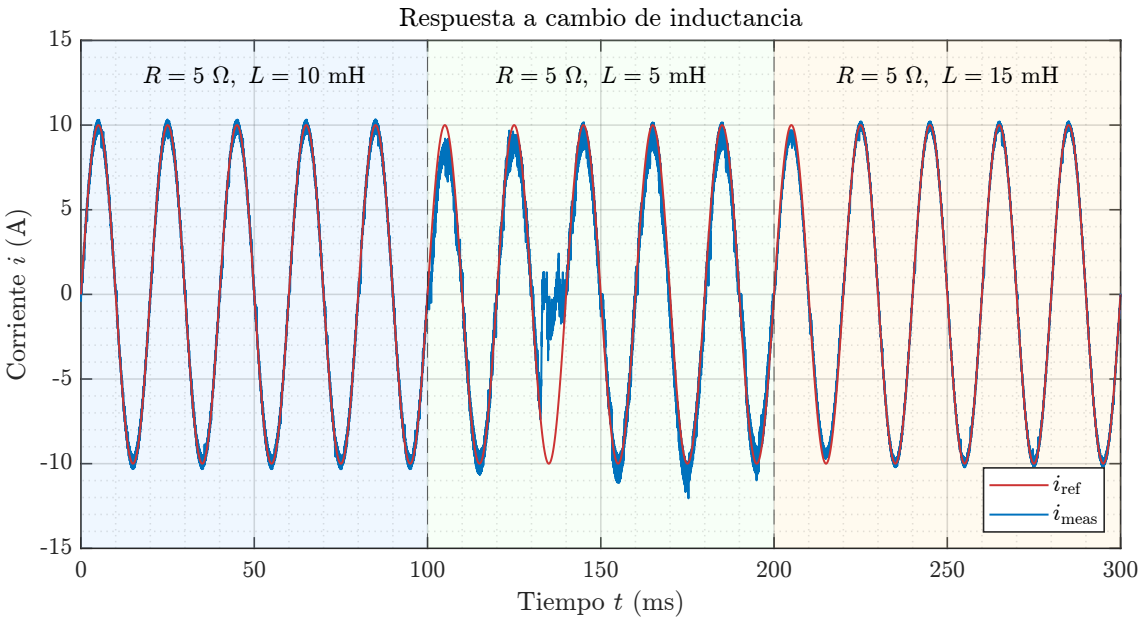


Figura 6.2: Respuesta del control Q-learning ante variaciones de la inductancia de carga en PLECS.

En la segunda ventana de tiempo, comparar el tiempo de estabilización observado con la respuesta natural del sistema estimado a partir de la constante de tiempo $\tau = L/R$, se aprecia una diferencia considerable. Para $L = 5$ mH y $R = 5\ \Omega$, el sistema físico debería alcanzar el régimen estacionario en torno a $4\tau \approx 4$ ms, sin embargo, el agente Q-learning requiere cerca de 50 ms para estabilizar la corriente. Esta diferencia se debe principalmente a la inestabilidad inicial que presenta el controlador durante el proceso de aprendizaje, donde la política aún no logra adaptarse completamente a la nueva dinámica del sistema.

Tabla 6.1: Métricas obtenidas para distintas variaciones de inductancia en el control Q-learning.

L [mH]	10	5	15
THD [%]	3.88	6.86	2.17
RMSE [%]	2.9	5.4	1.6

En la Tabla 6.1 se observa que tanto el THD como el RMSE de la corriente aumentan al reducir el valor de la inductancia, lo que indica una mayor sensibilidad del control ante dinámicas más rápidas del sistema.

6.1.3. Control de corriente frente a cambio frecuencia

Se realiza un experimento en el que se varía la frecuencia de una señal de referencia sinusoidal de 10 A de amplitud, con el objetivo de analizar su comportamiento al superar los límites operacionales. Para una corriente sinusoidal de 10 A, el límite físico teórico se encuentra en torno a los 300 Hz (Apéndice C.2).

En estado estacionario a 50 Hz, se incrementa la frecuencia sucesivamente a 100 Hz, 200 Hz, 300 Hz y 400 Hz. En la Figura 6.3 se observa que, para las frecuencias dentro del rango operativo previamente determinado, el control logra un adecuado seguimiento de la señal de referencia. Al superar el umbral de los 300 Hz, el agente pierde la capacidad de controlar la corriente de carga, lo cual se atribuye a que la fuente es incapaz de suministrar una corriente mayor debido a las limitaciones físicas impuestas por la tensión y la impedancia, según se describe en la Ecuación C.3.

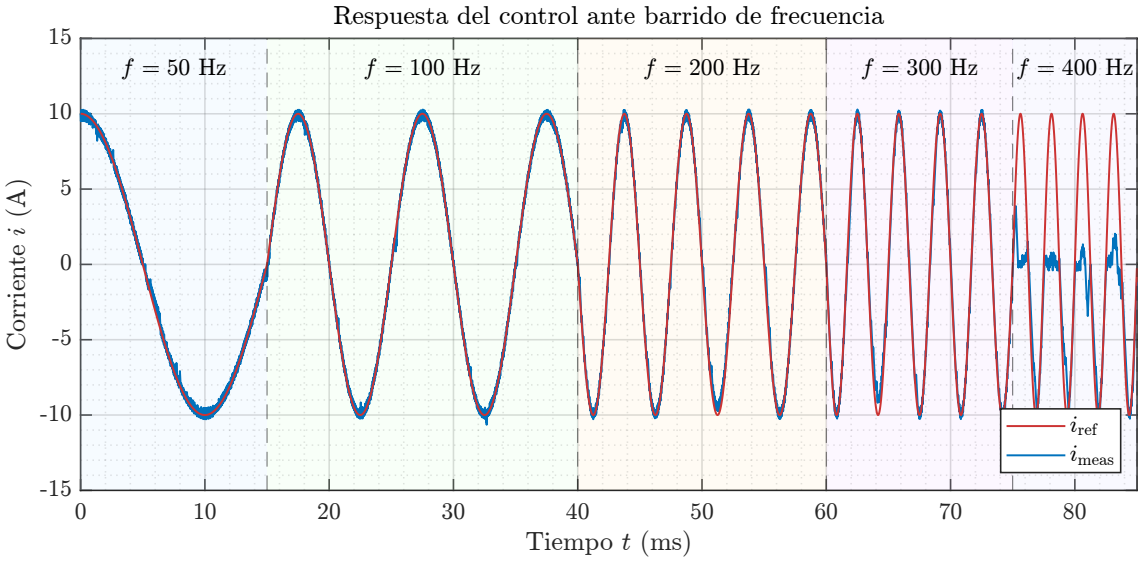


Figura 6.3: Respuesta cambio de frecuencia con $A_{ref} = 10$ A - agente QL en PLECS.

Este resultado evidencia que el desempeño del control Q-learning está condicionado por las restricciones físicas del sistema más que por su propia estabilidad. Dentro del rango admisible, el agente mantiene un seguimiento preciso, sin variaciones significativas en el error ni en el rizado, lo que demuestra un comportamiento robusto frente a cambios dinámicos de frecuencia. Al sobrepasar el límite teórico, la pérdida de control no se asocia a una falla en la política aprendida, sino al agotamiento de la capacidad de entrega de corriente de la fuente.

Tabla 6.2: Métricas del barrido en frecuencia para el control Q-learning.

f [Hz]	50	100	200	300	400
THD [%]	3.19	2.90	3.85	2.87	34.93
RMSE [%]	2.4	2.6	3.6	3.2	46.7

Se observa en la Tabla 6.2 que, hasta 300 Hz, las métricas de THD y RMSE se mantienen en niveles comparables entre sí, coherentes con el comportamiento observado en la Figura 6.3. En cambio, a 400 Hz ambas métricas aumentan de manera abrupta, lo que confirma la pérdida de capacidad de seguimiento al operar fuera del rango físicamente admisible.

6.1.4. Control de corriente frente a cambio de amplitud

Se realiza una variación en el valor de la amplitud de la corriente de referencia con el fin de verificar el desempeño del control dentro de la zona de operación.

En régimen estacionario, con una corriente de referencia de 5 A de amplitud y una frecuencia de 50 Hz, se efectúan cambios de amplitud a 15 A, 10 A y 20 A, manteniendo cuatro ciclos por cada nivel. Como se observa en la Figura 6.4, para todas las amplitudes exploradas el agente logra adaptarse y seguir correctamente la referencia.

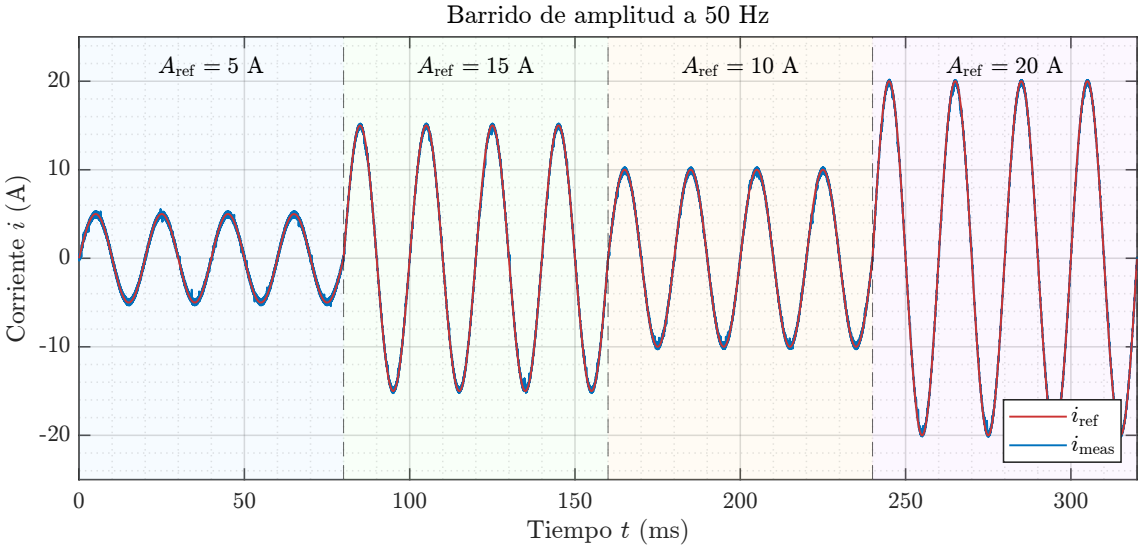


Figura 6.4: Respuesta cambio de amplitud a $f_{\text{ref}} = 50$ Hz - agente QL en PLECS.

A_{ref} [A]	5	15	10	20
THD [%]	6.62	2.10	2.59	1.75
RMSE [%]	4.8	1.7	2.6	1.4

Tabla 6.3: Métricas para distintas amplitudes de corriente en el control Q-learning.

En la Tabla 6.3 se presentan los valores de THD y RMSE para distintas amplitudes de la corriente de referencia. El RMSE se expresa como un porcentaje respecto a su respectiva amplitud de referencia (A_{ref}) con el objetivo de normalizar el error y permitir la comparación entre señales de diferente magnitud.

Se observa que, a medida que aumenta la amplitud de la referencia, el error relativo disminuye, evidenciando un mejor seguimiento del control para señales de mayor amplitud. En cambio, para la amplitud más baja (5 A), el RMSE relativo es mayor, lo que sugiere una mayor sensibilidad del agente a errores de cuantización o a la discretización del espacio de estados cuando las corrientes operan en rangos reducidos.

En todos los casos, los valores de THD se mantienen acotados y muestran que la calidad de la forma de onda no se deteriora de manera significativa al variar la amplitud de la corriente.

6.1.5. Control de corriente frente a referencia constante

En la Figura 6.5 se muestra el seguimiento de una referencia escalonada por parte del agente Q-learning para dos configuraciones del mecanismo Action Hold: AH=5 y AH=1. La señal de referencia presenta cuatro niveles de amplitud (5, 10, 15 y 20 A), permitiendo evaluar el comportamiento del agente en distintos rangos de operación.

Ambas configuraciones logran seguir la referencia, pero presentan diferencias claras en la calidad del seguimiento. Para AH=5, la señal de corriente medida exhibe una menor variabilidad y un comportamiento más estable, lo que se traduce en un error transitorio reducido y una oscilación acotada alrededor del valor deseado. Por el contrario, la configuración AH=1, al actualizar la acción en cada paso temporal, produce una señal más ruidosa y con mayor dispersión respecto de la referencia, particularmente visible en los escalones superiores donde la exigencia sobre el inversor es mayor.

En los tramos de mayor amplitud (15 A y 20 A), ambas señales tienden a subestimarse ligeramente respecto de la referencia, lo que indica que el agente enfrenta una mayor complejidad al controlar corrientes más elevadas. Sin embargo, incluso en este régimen, la configuración con AH=5 mantiene una trayectoria más consistente y con mejor alineación general respecto de la referencia.

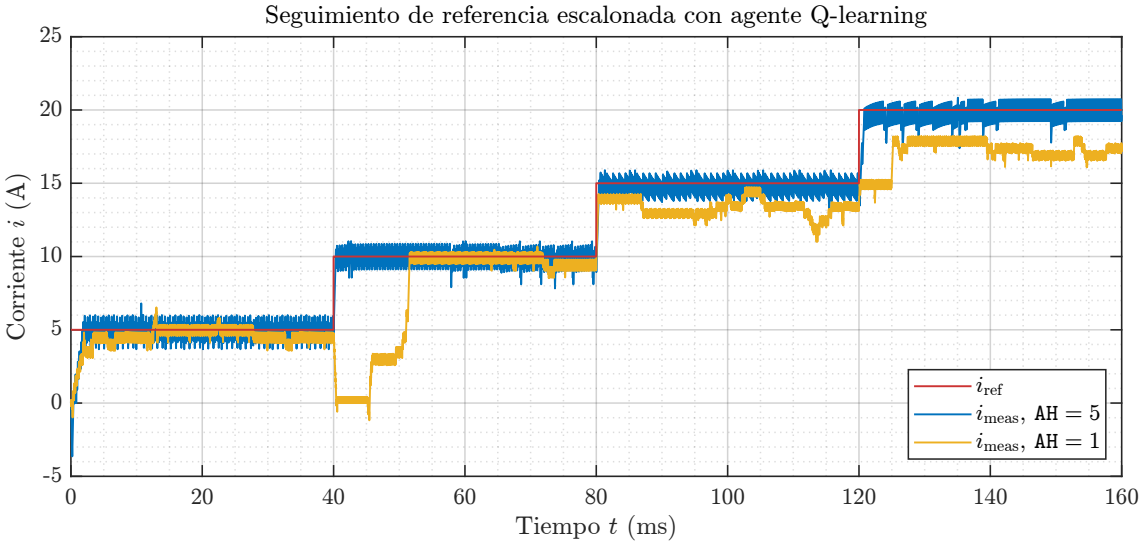


Figura 6.5: Respuesta a una señal escalón - agente QL en PLECS.

El análisis evidencia que el parámetro Action Hold influye de manera decisiva en el desempeño del agente Q-learning. La configuración AH=5 ofrece un seguimiento más estable y menos ruidoso, reduciendo la variabilidad de la corriente y manteniendo una trayectoria más próxima a la referencia en todos los escalones. En contraste, AH=1 incrementa la sensibilidad del agente, lo que conduce a una mayor dispersión y oscilación, afectando la calidad del seguimiento, especialmente en amplitudes altas.

Estos resultados indican que, para el problema de control de corriente considerado, un mayor Action Hold actúa como un mecanismo estabilizador beneficioso que mejora la respuesta práctica del sistema sin comprometer significativamente la precisión del seguimiento.

6.1.6. Efecto de Action Hold en control de señales cuadradas

Como se ha mencionado anteriormente, el parámetro AH tiene un impacto directo sobre las decisiones que toma el agente Q-learning durante el aprendizaje. Para señales suaves, como sinusoides, al utilizar valores más pequeños de AH, las señales presentan menores valores de RMSE y THD, dada la mayor suavidad que se obtiene al decidir con mayor frecuencia las acciones a ejecutar. Sin embargo, esto implica un mayor costo computacional, ya que la matriz Q debe actualizarse con mayor frecuencia. Por el contrario, al aumentar el valor de dicho parámetro, las métricas mencionadas tienden a incrementarse, debido a que se permite que la dinámica del sistema evolucione durante más tiempo entre decisiones, generando corrientes menos suaves en la carga, aunque con un menor costo de cómputo y una frecuencia de conmutación más baja.

En la Figura 6.6 se muestra el caso de una señal de referencia cuadrada, donde se compara el aprendizaje de un agente para valores de AH iguales a 1 y 5. En otras palabras, en la primera opción el agente toma decisiones cada 10 μs, mientras que la segunda lo hace cada 50 μs. Se puede observar que para AH = 1, el control no logra adaptarse al seguimiento de la señal cuadrada, mientras que el segundo lo consigue aproximadamente a los 40 ms, luego de dos ciclos de aprendizaje.

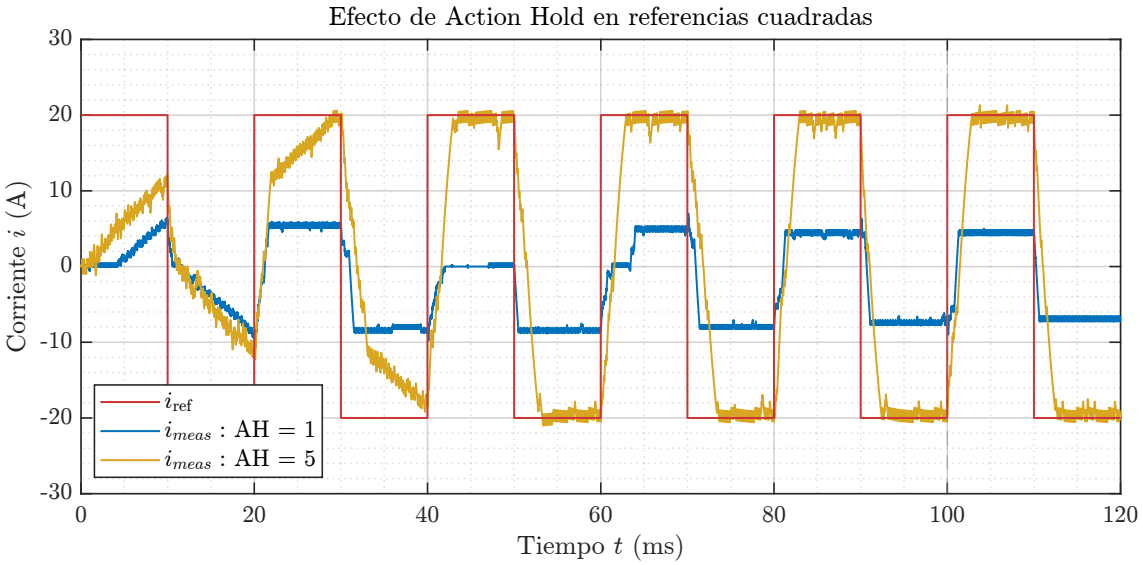


Figura 6.6: Efecto de parámetro Action Hold en la respuesta a señal cuadrada - agente QL en PLECS.

Estos resultados evidencian que un valor de Action Hold demasiado bajo dificulta que el agente aprenda un comportamiento estable, mientras que un valor moderado facilita la captura de la dinámica del sistema y favorece una mejor generalización de los estados.

6.1.7. Ajustes alternativos para el control Q-learning

Como se pudo apreciar en las distintas pruebas realizadas en PLECS, el agente logra efectuar un seguimiento adecuado de diversas referencias aun cuando el sistema se ve sometido a perturbaciones. Junto con los hiperparámetros seleccionados para Q-learning (apartado 5.2), se mostró en la sección anterior que el parámetro Action Hold tiene una relevancia crucial para un aprendizaje correcto del agente. No obstante, este parámetro presenta ciertas limitaciones cuando se extiende el tiempo total de simulación.

A continuación, se consideran los casos más críticos de la sección anterior para analizar el comportamiento del control en tramos más prolongados de simulación: el cambio de inductancia de 10 mH a 5 mH y corrientes de referencia cuadradas de gran amplitud.

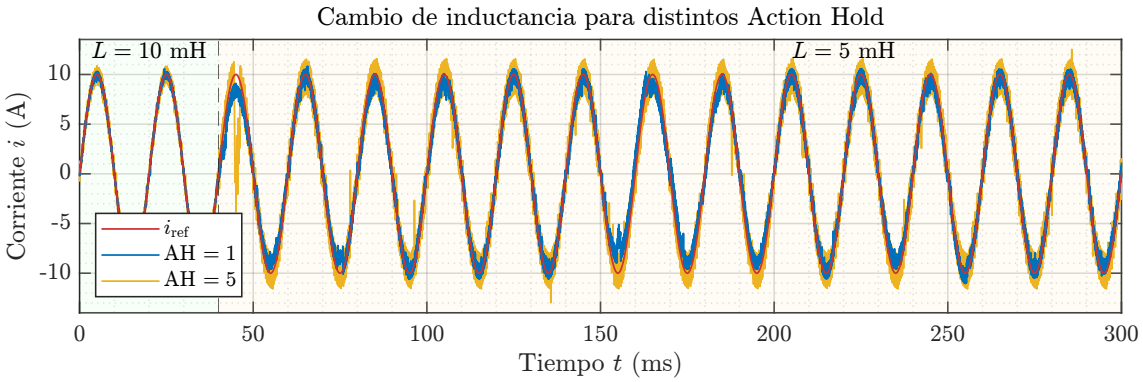


Figura 6.7: Efecto de parámetro Action Hold en la respuesta a cambio de inductancia.

En la Figura 6.7 se muestra la respuesta del control de corriente ante el cambio de inductancia a la mitad de su valor original. La corriente de referencia se mantiene fija en 10 A y 50 Hz. La carga conserva una inductancia de 10 mH durante el primer segundo de simulación, para luego disminuir instantáneamente a 5 mH.

En color azul se presenta la respuesta obtenida con un Action Hold (AH) de 1, la cual evidencia dificultades para alcanzar el valor máximo de corriente en varios ciclos. En contraste, la respuesta con AH $= 5$ (línea amarilla) presenta un comportamiento más estable en amplitud, aunque con un rizado más visible en su forma de onda, producto de la menor frecuencia de actualización del control.

Action Hold	A_1 [A]	THD [%]
1	9.450	0.501
5	9.731	0.354

Tabla 6.4: Comparación de distorsión para cambio de inductancia $f_1 = 50$ Hz.

En la Tabla 6.4 se muestra la comparación cuantitativa entre ambas señales. Se observa que entre AH $= 1$ y AH $= 5$ la distorsión armónica aumenta en torno a un 42 %. Este efecto se traduce en una reducción de la amplitud de la componente fundamental de la corriente con AH $= 1$.

En la Figura 6.8 se aprecia con mayor claridad el impacto del parámetro AH. La referencia consiste en una señal sinusoidal de 20 A y 50 Hz durante el primer segundo, que luego cambia instantáneamente a una señal cuadrada de igual amplitud y frecuencia.

El control con AH = 1 intenta sin adaptarse inicialmente a la nueva forma de onda, pero pierde gradualmente su capacidad de seguimiento tras los primeros ciclos. En cambio, con AH = 5, el agente sigue correctamente la señal cuadrada durante aproximadamente ocho ciclos consecutivos. A partir de ese punto, comienzan a presentarse fallas esporádicas en los flancos, donde la corriente queda en cero durante medio ciclo para luego recuperar el seguimiento en el siguiente. Este comportamiento se repite de manera intermitente a lo largo de la simulación.

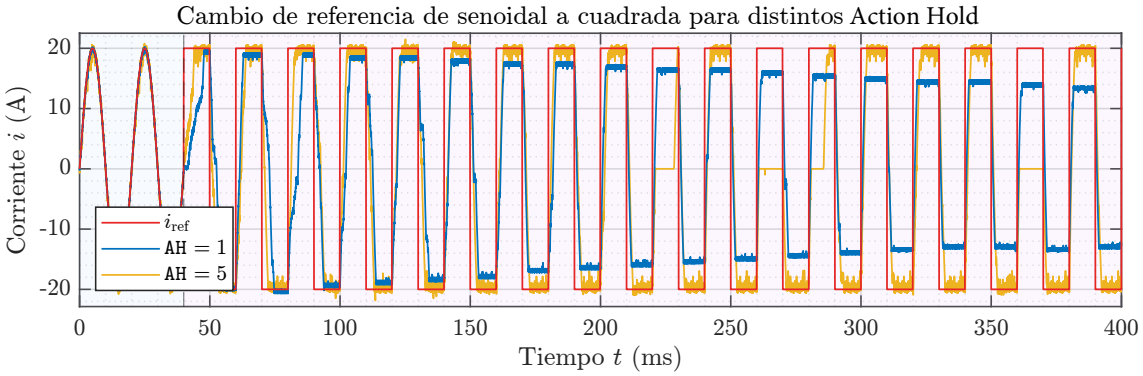


Figura 6.8: Efecto del parámetro Action Hold ante un cambio de referencia sinusoidal a cuadrada.

En la siguiente sección se propone incorporar técnicas adicionales al agente Q-learning para mejorar su capacidad de seguimiento ante señales con transiciones bruscas, manteniendo la estabilidad del control.

6.1.8. Evaluación de mecanismos estabilizadores de control

Tal como se detalla en el Apéndice A, el agente Q-learning puede complementarse con ciertos mecanismos destinados a suavizar su política de acción y reducir comportamientos no deseados durante la conmutación. En dicho apéndice se describen tres alternativas: *Action Hold* (AH), *Action Sticky* (AS) y el umbral Q_{thresh} (QT), todas orientadas a mejorar la estabilidad del control en escenarios donde la toma de decisiones discreta interactúa con una dinámica eléctrica rápida.

En esta sección se profundiza en la experimentación realizada en *PLECS Standalone* al incorporar los dos mecanismos restantes, AS y QT, con el fin de observar su efecto directo sobre el comportamiento del inversor y la política del agente.

En primera instancia se evalúa cambio de inductancia en la carga, cuyo objetivo es analizar la estabilidad del control y su capacidad de seguimiento frente a una variación abrupta en la dinámica del sistema. En este caso, la corriente de referencia se mantiene sinusoidal con amplitud de 10 A y frecuencia de 50 Hz, mientras que la inductancia de la carga se reduce a la mitad en el instante $t = 1$ s, pasando de 10 mH a 5 mH.

El segundo experimento corresponde a un cambio en la forma de la señal de referencia, manteniendo fija la carga en $5\ \Omega$–10 mH. La señal de referencia es inicialmente una onda sinusoidal de 20 A y 50 Hz durante el primer segundo, y luego cambia instantáneamente a una señal cuadrada de igual amplitud y frecuencia. Este escenario busca evaluar la capacidad del agente para adaptarse a referencias con transiciones bruscas y alta pendiente. Dado que en una señal cuadrada de amplitud 20 A el valor eficaz es también 20 A, se utiliza el valor RMS de la corriente medida como indicador adicional de desempeño, ya que permite identificar qué combinación de parámetros produce una respuesta más cercana a la forma ideal de la referencia.

Con el fin de evaluar el efecto combinado de los mecanismos de decisión propuestos, se analizaron distintas configuraciones del control Q-learning ajustando los parámetros Action Hold (AH), *Action Sticky* (AS) y el umbral de diferencia en el valor Q (Q_{thresh}, QT). Estas combinaciones permiten estudiar la influencia de la frecuencia de actualización de la acción, la persistencia temporal en la política y la sensibilidad del agente ante mejoras marginales en el valor Q. En los resultados, cada configuración se representa mediante el formato *AH-AS-QT*, lo que facilita la

comparación directa del desempeño de cada conjunto de parámetros frente a cambios en la carga o en la referencia.

AH-AS-QT	**10 mH (0–1 s)**			**5 mH (1–2 s)**			i_{RMS} [A]
	RMSE [%]	f_{sw} [kHz]	THD [%]	RMSE [%]	f_{sw} [kHz]	THD [%]	(cuadrada)
1-0.0-0.0	6.239	30.25	0.353	7.385	30.41	0.501	10.91
5-0.0-0.0	5.390	14.23	0.381	10.031	14.71	0.354	17.44
1-0.0-0.1	5.714	29.95	0.792	5.996	31.03	0.313	11.04
1-0.0-0.5	6.151	12.10	0.415	5.362	25.68	0.144	11.28
1-0.1-0.0	6.580	26.77	0.494	7.170	28.93	0.316	17.06
1-0.1-0.1	7.199	27.36	0.424	6.563	28.78	0.435	17.45
1-0.1-0.5	7.003	11.61	0.506	6.562	20.76	0.256	17.78
1-0.2-0.0	6.703	25.06	0.489	9.771	26.09	0.504	17.96
1-0.2-0.1	7.791	24.46	0.348	6.837	25.99	0.203	17.97
1-0.2-0.5	8.172	11.13	0.767	7.133	19.38	0.212	18.03
1-0.3-0.0	9.025	22.96	0.524	8.824	24.05	0.768	17.97
1-0.3-0.1	7.504	21.78	0.878	7.714	23.39	0.592	17.99
1-0.3-0.5	8.958	10.41	0.772	7.948	17.77	0.184	18.14

Tabla 6.5: Resumen de métricas por configuración para los tramos de 10 mH y 5 mH, incluyendo el espacio para registrar el valor RMS de la corriente medida bajo referencia cuadrada.

A partir de los resultados mostrados en la Tabla 6.5, se puede observar que la incorporación de los parámetros AS y QT permite mejorar el comportamiento general del control Q-learning frente a variaciones abruptas en la carga y en la referencia. En particular, las configuraciones con QT distinto de cero muestran una respuesta más estable y con menor distorsión armónica total, lo que sugiere que el umbral sobre el valor Q ayuda a filtrar decisiones poco significativas del agente.

Si bien aumentar el Action Hold reduce la frecuencia de conmutación, también limita la capacidad de respuesta del sistema ante cambios bruscos, lo que se refleja en un aumento del error de seguimiento al reducir la inductancia. En contraste, combinar un AH corto con una probabilidad moderada de acción persistente (AS) y un umbral de histéresis sobre el valor Q (QT) logra un mejor equilibrio entre estabilidad y precisión.

De todas las configuraciones, la combinación 1-0.3-0.5 presenta el desempeño más robusto. Esta logra mantener una frecuencia de conmutación moderada, un bajo nivel de distorsión (THD) y una corriente eficaz cercana al valor ideal en el caso de referencia cuadrada. En comparación, 1-0.2-0.5 también ofrece buenos resultados, especialmente en términos de calidad espectral, aunque con un leve aumento del error RMS.

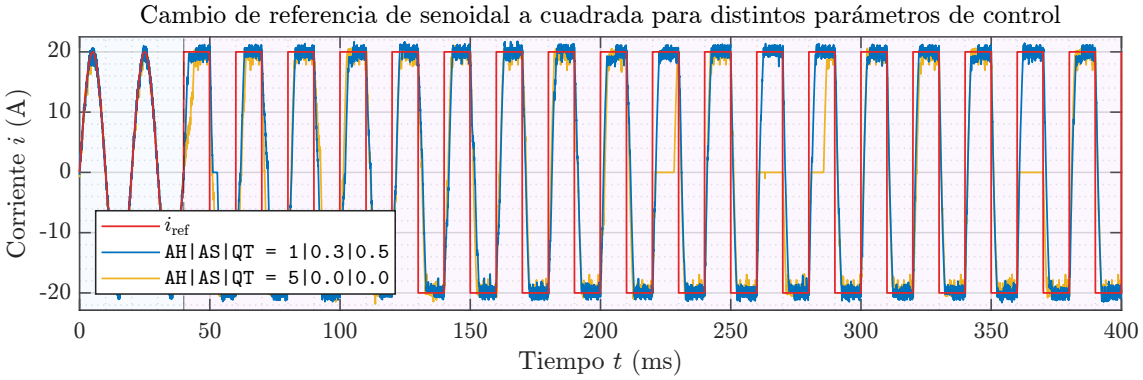


Figura 6.9: Efecto de parámetros de control frente a un cambio de referencia sinusoidal a cuadrada

En la Figura 6.9 se muestra la comparación entre un agente Q-learning con un AH = 5 puro y otro que incorpora los parámetros AS y QT. A diferencia de lo observado con AH = 1, ambas configuraciones logran mantener el seguimiento de una señal cuadrada de gran amplitud a lo largo del tiempo. Sin embargo, al incorporar los nuevos parámetros, los flancos que antes el agente dejaba de seguir (línea amarilla en la figura) desaparecen, dando lugar a un seguimiento más consistente y estable.

En síntesis, los resultados confirman que incorporar correlación temporal en la política del agente y un umbral de decisión en la actualización de Q mejora significativamente la estabilidad del control, sin comprometer la calidad de seguimiento ni la eficiencia de conmutación.

6.1.9. Respuesta de control frente a distintas función de recompensas

Con el objetivo de analizar la sensibilidad del agente Q-learning frente a la definición de la función de recompensa, se implementaron distintas formulaciones que ponderan de manera diferente el error de seguimiento y el esfuerzo de control. Todas las variantes fueron evaluadas utilizando las métricas definidas en la Sección 5.1, con el fin de comparar de manera consistente la calidad de seguimiento, la calidad espectral y el esfuerzo de conmutación asociado a cada formulación de recompensa.

Para ello, se definen las siguientes funciones de recompensas:

- Recompensa cuadrática negativa del error:

$$r_1(t) = -\big(i_{\text{ref}}(t) - i_{\text{meas}}(t)\big)^2 \tag{6.1}$$

- Recompensa normalizada en función del error relativo:

$$r_2(t) = -|i_{\text{ref}}(t) - i_{\text{meas}}(t)| \tag{6.2}$$

- Recompensa penalizada por conmutaciones o equivalentemente cambios de acción Δa:

$$r_3(t) = -\big(i_{\text{ref}}(t) - i_{\text{meas}}(t)\big)^2 - \lambda|\Delta a| \tag{6.3}$$

  donde $\Delta a = \{0, 1\}$ es cero si $a_{t-1} = a_t$, y uno en caso contrario. λ es un pequeño factor de penalización.

- Recompensa exponencial del error:

$$r_4(t) = \exp(-k_{\text{exp}} \cdot |i_{\text{ref}}(t) - i_{\text{meas}}(t)|) \tag{6.4}$$

Para cada variante de recompensa se evaluó el desempeño del agente en un escenario de seguimiento de una referencia senoidal de 50 Hz y amplitud de 15 A. En todos los casos se utilizó un *action hold* AH = 1, una probabilidad de *action stickiness* AS = 0.3 y un umbral en la función de valor-acción QS = 0.5. La única excepción corresponde a la recompensa exponencial, para la cual se deshabilitó el mecanismo de histéresis debido a que, dadas las características de discretización de la tabla Q, dicha combinación condujo a inestabilidad en el proceso de aprendizaje. La Tabla 6.6 resume los resultados obtenidos en estado estacionario para cada función de recompensa, considerando las métricas de ganancia relativa, RMSE normalizado, distorsión armónica total y frecuencia de conmutación promedio. Finalmente, en la Figura 6.10 se presentan las formas de

onda comparativas, las cuales permiten analizar el comportamiento dinámico inicial y la rapidez de convergencia de cada agente.

Función de recompensa	G	THD[%]	RMSE [%]	f_{sw} [kHz]
r_1 (6.1)	0.994	6.20	4.51	5.30
r_2 (6.2)	0.997	5.66	4.00	6.35
r_3 (6.3)	0.996	4.75	3.35	7.65
r_4 (6.4)	0.990	2.15	1.68	15.35

Tabla 6.6: Métricas en estado estacionario para diferentes funciones de recompensa.

Las tres primeras funciones de recompensa (Ecuaciones 6.1 - 6.3), presentan respuestas transitorias rápidas, alcanzando un seguimiento adecuado durante los primeros ciclos de la señal. Esto se asocia a que dichas recompensas mantienen una sensibilidad relativamente alta frente a errores moderados, promoviendo correcciones tempranas de la acción del agente.

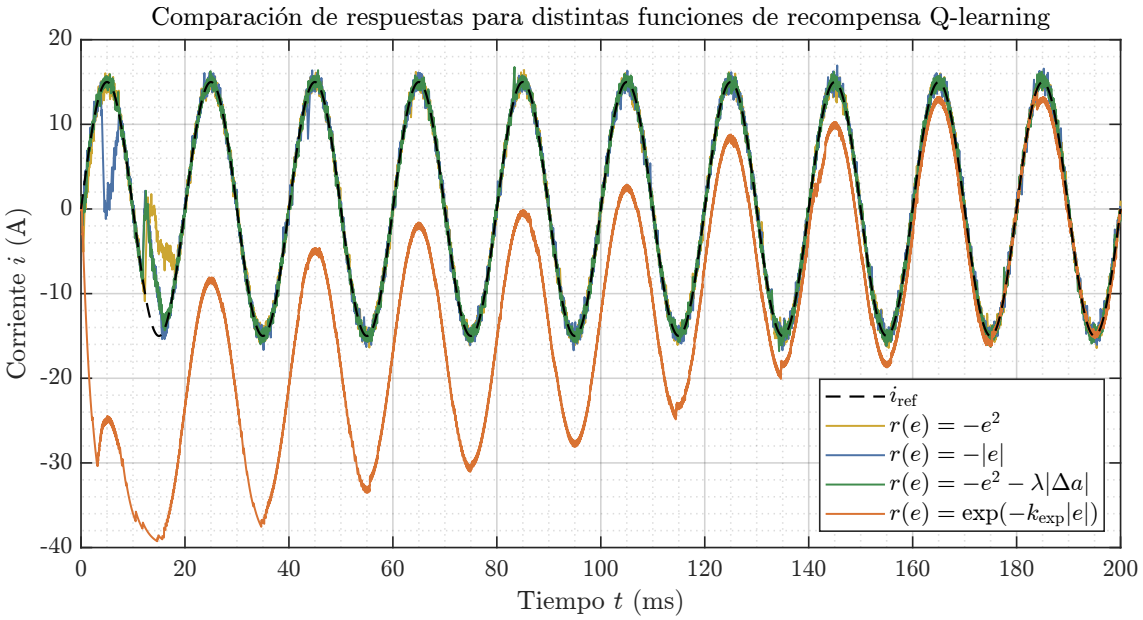


Figura 6.10: Respuesta de control de corriente con agente Q-learning para distintas recompensas.

En contraste, la recompensa exponencial exhibe un comportamiento significativamente más lento, requiriendo aproximadamente 190 ms para lograr un seguimiento satisfactorio. Este retardo se explica por la propia forma de la función exponencial: para valores de error no despreciables, el gradiente del término $\exp(-k|e|)$ es reducido, lo que disminuye la presión del agente por corregir el error y retrasa la convergencia de la política.

En estado estacionario, la recompensa exponencial (Ecuación 6.4) alcanza las mejores métricas de precisión, logrando simultáneamente el menor RMSE y la menor distorsión armónica total. No obstante, estas mejoras se obtienen a costa de un incremento significativo en la frecuencia de conmutación, que se vuelve considerablemente más alta que en el resto de las variantes. Por otro lado, las recompensas cuadrática y absoluta (Ecuaciones 6.1 y 6.2) presentan un equilibrio más favorable entre calidad de seguimiento y esfuerzo de conmutación, aunque con un THD mayor. En cuanto a la recompensa penalizada por cambios de acción (Ecuación 6.3), si bien estaba diseñada para reducir el switching, los resultados en la Tabla 6.6 muestran que este término puede inducir un aumento inesperado de la frecuencia de conmutación, producto de la interacción entre

la discretización del estado y la estructura de la recompensa, que incentiva ajustes más frecuentes ante variaciones pequeñas del error.

6.1.10. Impacto de la dimensión de tabla Q

En un principio se propuso una tabla que cubre una discretización del error entre -40 A y 40 A con un paso de 0.5 A. Como se mostró en la sección 3.3, con un número de estados $N_e = 161$ y un número de acciones $N_a = 3$ se obtiene una tabla Q de dimensión Dim(Q) = $N_e \times N_a = 483$. Sin embargo, dado que el objetivo del agente es reducir el error, la política siempre tiende a seleccionar acciones que aproximan el sistema a valores cercanos a cero. Esto provoca que una parte importante de la tabla Q correspondiente a errores de gran magnitud se utilice en muy pocas ocasiones, lo que constituye un uso poco eficiente del método, ya que una tabla más extensa requiere más tiempo para que el agente aprenda la política adecuada.

En la Figura 6.11 se muestra el desempeño del control de corriente durante 500 ms utilizando una tabla Q inicializada en cero. Se modifica la amplitud de la corriente con el propósito de observar la evolución del error durante el proceso de control. Se aprecia que, excepto durante los primeros 50 ms en los que el agente aún se encuentra en etapa de aprendizaje, el error se mantiene la mayor parte del tiempo alrededor de cero con un valor de RMSE igual a 0.76 A. Este resultado sugiere que el espacio de estados discretizado a partir del error presenta una marcada concentración en valores pequeños, lo que respalda la idea de ajustar la discretización para aprovechar mejor las regiones de mayor relevancia para la política óptima.

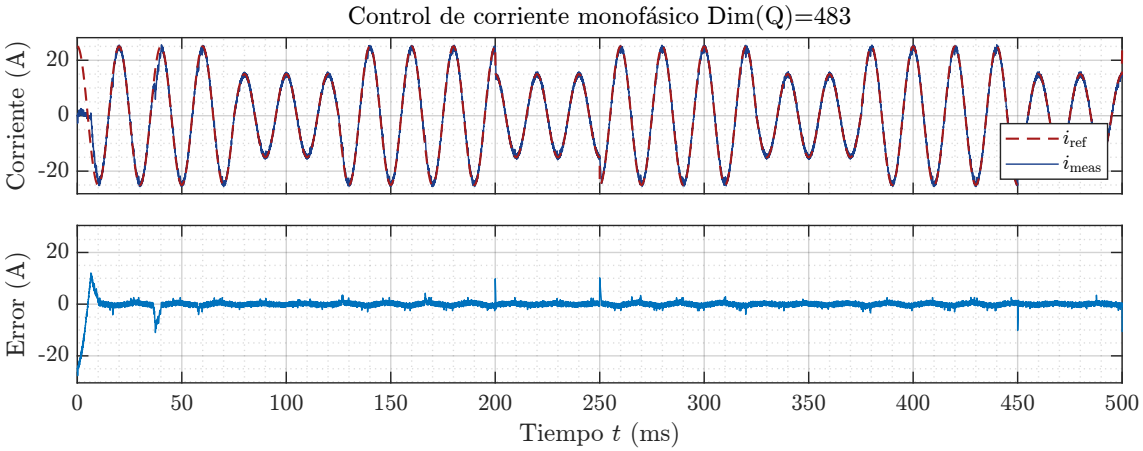


Figura 6.11: Control de corriente con agente Q-learning con Dim(Q) = 483.

En la Ecuación 3.10 del Apartado 3.3, utilizada para generar los estados discretos del control mediante un agente Q-learning, los valores de error que exceden los límites mínimo y máximo quedan acotados dentro del rango definido. En estas situaciones el agente aplica la acción asociada al estado saturado, independiente de la magnitud real del error. Este comportamiento es análogo al efecto de saturación de la señal de control en un regulador PI, ya que la salida permanece limitada una vez que el error supera cierto umbral. Esta observación permite reducir el tamaño de la tabla Q y disminuir el costo computacional del control sin comprometer el objetivo, debido a que los estados extremos no entregan información relevante para la política óptima.

Considerando un espacio de estados con $e_{\text{min}} = -10$ A, $e_{\text{max}} = 10$ A y e_{step} = 0.5 A, y aplicando la Ecuación 3.9, se obtiene un número de estados igual a $N_e = 41$. Si se mantiene el número de

acciones en $N_a = 3$, la dimensión de la tabla Q resulta en $\text{Dim}(Q) = 123$. Esta reducción equivale a una disminución del 74.5 % respecto del tamaño original de la tabla Q.

En la Figura 6.12 se muestra el control de corriente bajo las mismas condiciones, modificando únicamente la dimensión de la tabla Q. Se observa que, frente a la misma referencia, el control no pierde estabilidad y además presenta una mayor rapidez de aprendizaje, ya que no aparece el pico de error asociado al proceso de adaptación alrededor de los 40 ms que se aprecia en la Figura 6.11. Considerando el intervalo posterior a los 50 ms, el valor de RMSE es igual a 0.74 A, lo que representa una ligera mejora en comparación con el tamaño de tabla original.

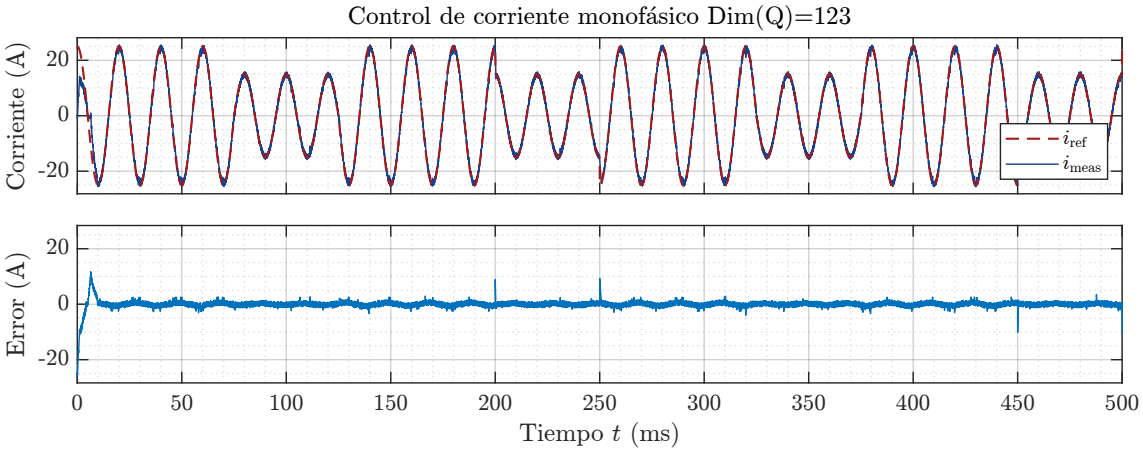


Figura 6.12: Control de corriente con agente Q-learning con Dim(Q) = 123.

Los resultados obtenidos permiten concluir que la discretización original del error, que considera un rango amplio y un número elevado de estados, no aporta beneficios significativos al desempeño del control basado en Q-learning. El análisis de las simulaciones muestra que el agente opera la mayor parte del tiempo en un entorno cercano al error cero, por lo que una fracción considerable de la tabla Q queda prácticamente sin uso y no contribuye a la construcción de la política óptima.

Al adoptar una discretización más acotada, el tamaño de la tabla Q disminuye en un $74{,}5\,\%$. Aunque en ambos casos se trata de un espacio de estados reducido y manejable, a medida que aumenta la complejidad del problema y se incorporan más variables a discretizar, esta reducción puede representar una diferencia significativa en el costo computacional del control. Además, los resultados muestran que el agente aprende con mayor rapidez y alcanza un error cuadrático medio ligeramente inferior al obtenido con la tabla original, lo que confirma que el desempeño del control no solo se mantiene, sino que presenta una mejora marginal.

6.2. Control de corriente monofásico con agente Deep Q-Network

De manera análoga al caso Q-learning, a continuación se presentan pruebas de desempeño del control de corriente implementado mediante un agente DQN, entrenado en MATLAB y exportado a PLECS a través de una *Look-up Table* (LUT). Los hiperparámetros seleccionados de las pruebas realizadas en MATLAB se muestran en la Tabla 6.7.

Tabla 6.7: Hiperparámetros considerados para la exportación de la LUT.

N_n	α	γ	$\varepsilon_{\text{decay}}$	TSF	$\lvert\mathcal{M}\rvert$
32	5e-4	0.99	1.5e-5	1e-4	256

La LUT importada corresponde a una matriz que indica la acción que debe ejecutar el agente, donde las columnas representan los valores de corriente medida y las filas los valores de corriente de referencia. Dichas corrientes son discretizadas durante la exportación, lo que permite definir diferentes resoluciones para el espacio de estados. En primera instancia, se realizaron pruebas con una LUT de resolución de 100 mA, la cual entrega una respuesta suave ante referencias sinusoidales.

6.2.1. Control de corriente frente a cambio de resistencia

Se realizan cambios instantáneos en el valor de la resistencia de carga para analizar la respuesta del control frente a este tipo de perturbaciones. El experimento consiste en alcanzar primero un estado estacionario con una corriente de referencia de 10 A de amplitud y 50 Hz de frecuencia. Luego, en un instante determinado, se produce una disminución abrupta de la resistencia desde 5 Ω a 2 Ω (un 40 % de su valor inicial), y posteriormente se incrementa a 10 Ω (200 % del valor inicial). Como se observa en la Figura 6.13, durante todas las ventanas de tiempo el control logra seguir adecuadamente la referencia de corriente.

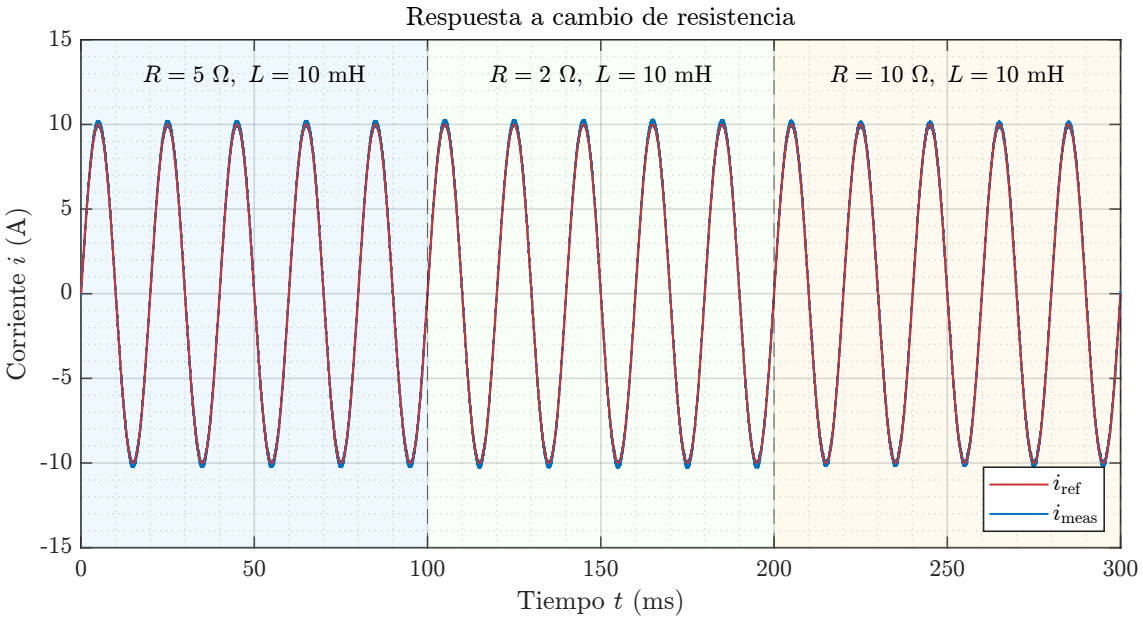


Figura 6.13: Respuesta cambio de resistencia de carga - agente DQN en PLECS.

Cualitativamente, el seguimiento de corriente obtenido con el agente DQN es más suave y continuo en todos los tramos respecto al Q-learning (Figura 6.1). Cuantitativamente, la distorsión armónica total se reduce de 3.33 % a 1.31 %, reflejando la mejora observada en la forma de onda.

6.2.2. Control de corriente frente a cambio de inductancia

Al igual que en el experimento con cambio de resistencia, se realizan variaciones instantáneas en el valor de la inductancia para analizar la respuesta del control frente a este tipo de perturbaciones. Luego de alcanzar el estado estacionario con una corriente de referencia de 10 A de amplitud y 50 Hz de frecuencia, se produce una disminución abrupta de la inductancia desde 10 mH a 5 mH (un 50 % de su valor inicial) y, posteriormente, se incrementa a 15 mH (150 % del valor inicial).

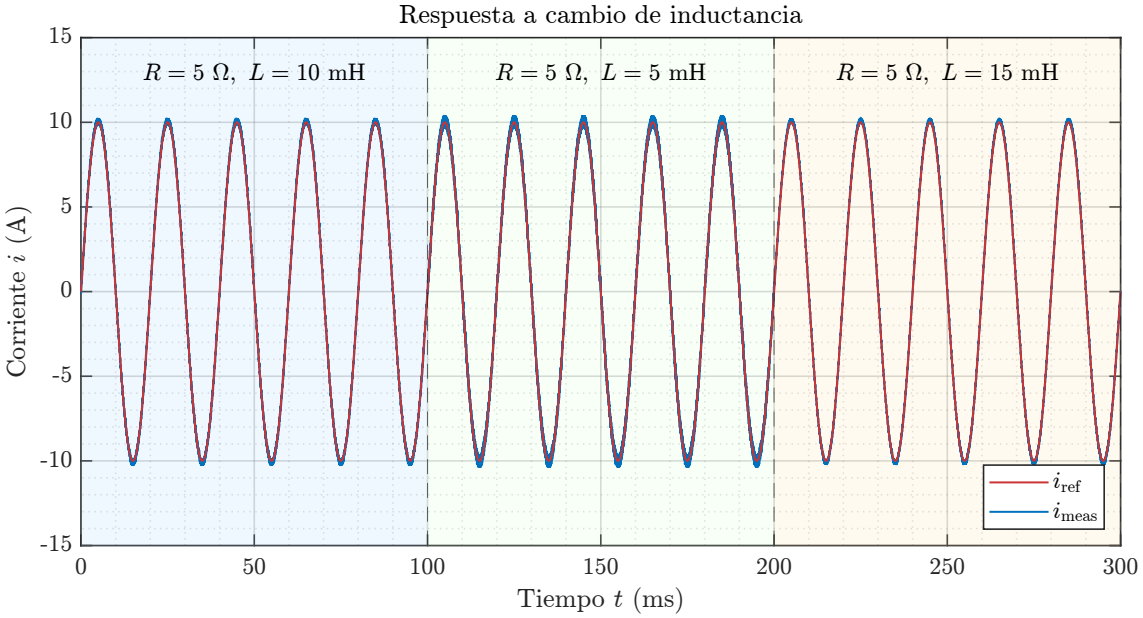


Figura 6.14: Respuesta cambio de inductancia de carga - agente DQN en PLECS.

A diferencia del control Q-learning, se observa que el agente DQN presenta un mayor rizado al reducir la inductancia a la mitad, el cual vuelve a disminuir al aumentar dicho valor. Este comportamiento se explica porque, al disminuir la inductancia, aumenta la rapidez natural del sistema, lo que produce mayores variaciones instantáneas de corriente. En comparación con el control Q-learning, el agente DQN no muestra problemas de adaptación inicial (Figura 6.2), aún habiendo sido entrenado con la carga original, mientras que el Q-learning debe ajustar su política *online* para adaptarse a la nueva dinámica impuesta por la constante de tiempo de la carga RL.

L [mH]	10	5	15
THD [%]	1.31	2.23	0.97
RMSE [%]	1.19	1.97	1.26

Tabla 6.8: Métricas obtenidas para distintas variaciones de inductancia en el control Deep Q-Network.

En la Tabla 6.8 se presentan métricas del control DQN bajo las variaciones de inductancia. En concordancia con la dinámica del sistema RL, el THD tiende a disminuir al aumentar L. En comparación con el caso Q-learning, el DQN presenta valores sistemáticamente menores de THD

y RMSE en los tres escenarios, lo que se refleja en un seguimiento más suave y con menor error relativo.

6.2.3. Control de corriente frente a cambio de frecuencia

Al igual que en Q-learning, límitación física del inversor (Apéndice C.2), se realiza un experimento en el que se varía la frecuencia de una señal de referencia sinusoidal de 10 A de amplitud, con el objetivo de analizar su comportamiento al superar los límites operacionales. Para una corriente sinusoidal de 10 A, el límite físico se encuentra en torno a los 300 Hz.

En estado estacionario a 50 Hz, la frecuencia se incrementa sucesivamente a 100 Hz, 200 Hz, 300 Hz y 400 Hz. En la Figura 6.15 se observa que, para las frecuencias dentro del rango operativo previamente determinado, el control logra un seguimiento adecuado de la señal de referencia. A 400 Hz, a diferencia del control Q-learning que no lograba adaptarse y generaba una forma de onda errática, el agente DQN mantiene el intento de seguir la sinusoide, aunque comienza a salir de su zona de operación.

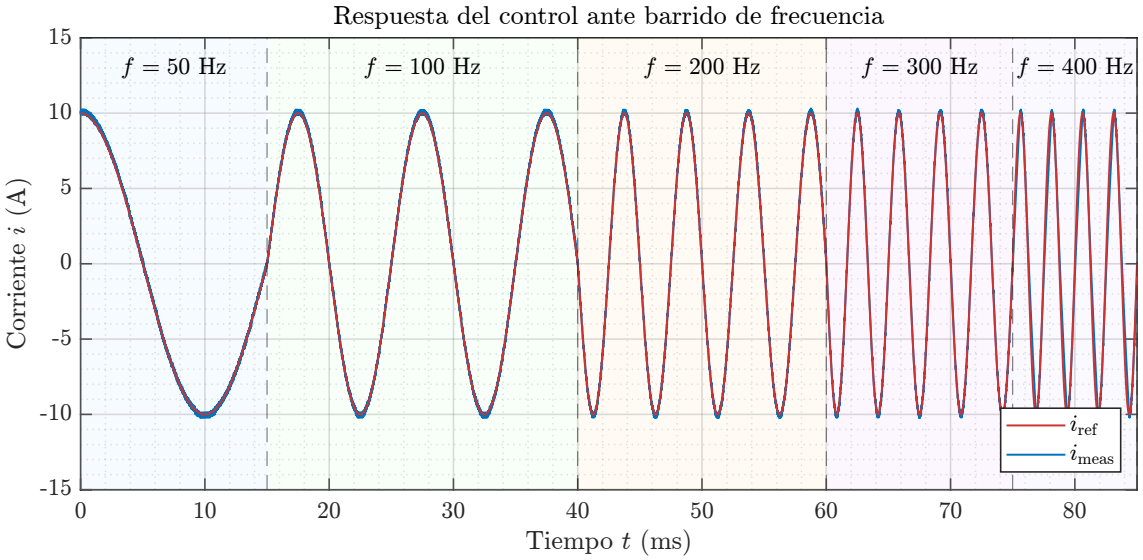


Figura 6.15: Respuesta cambio de frecuencia con $A_{ref} = 10$ A - agente DQN en PLECS.

En términos comparativos, el agente DQN conserva una respuesta más estructurada a 400 Hz que el caso Q-learning, manteniendo una forma aproximadamente sinusoidal aunque con mayor atenuación y distorsión, lo que es coherente con la cercanía al límite físico impuesto por la relación tensión–impedancia (Apéndice C.2).

En términos de degradación progresiva, los resultados muestran que el desempeño se mantiene prácticamente constante hasta 300 Hz, mientras que a 400 Hz se observa un aumento notorio tanto en THD como en RMSE (Tabla 6.9), consistente con la aproximación al límite físico de operación del inversor (Apéndice C.2). En consecuencia, el rango en el cual el seguimiento permanece comparable al caso base se extiende más allá de 300 Hz, aunque comienza a deteriorarse de forma apreciable al acercarse a 400 Hz.

Tabla 6.9: Métricas del barrido en frecuencia para el control Deep Q-Network.

f [Hz]	50	100	200	300	400
THD [%]	1.32	1.34	1.40	1.40	6.70
RMSE [%]	1.20	1.26	1.50	1.91	11.50

En la Tabla 6.9 se observa que, para las frecuencias dentro de la zona de operación, las métricas se mantienen por debajo del 2 %, lo que evidencia un seguimiento suave y con bajo error.

6.2.4. Control de corriente frente a cambio de amplitud

Al igual que en el caso anterior, se realiza una variación en la amplitud de la corriente de referencia con el fin de analizar el comportamiento del controlador ante variaciones del nivel de referencia dentro de un rango representativo de operación.

En régimen estacionario, con una corriente de referencia de 5 A y una frecuencia de 50 Hz, se efectúan cambios de amplitud a 15 A, 10 A y 20 A, manteniendo cuatro ciclos por cada nivel. Como se observa en la Figura 6.16, para todas las amplitudes exploradas el agente logra adaptarse y seguir correctamente la señal de referencia.

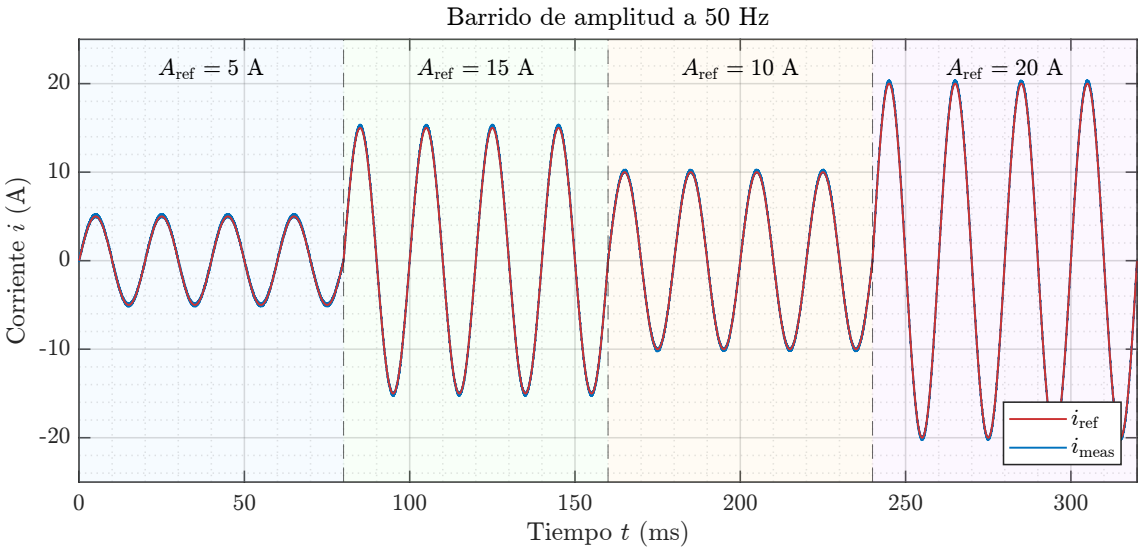


Figura 6.16: Respuesta ante cambios de amplitud con $f_{\text{ref}} = 50$ Hz — agente DQN en PLECS.

En la Tabla 6.10 se aprecia que, para las distintas amplitudes evaluadas, las métricas de THD y RMSE se mantienen bajas y con variaciones acotadas. Al comparar con los resultados obtenidos mediante Q-learning en la Tabla 6.3, se observa un comportamiento similar entre amplitudes equivalentes, reduciendo drásticamente la distorsión armónica hasta en un 60 % en el caso de una amplitud de 5 A. Además, se evidencia una disminución de la distorsión a medida que aumenta la corriente de referencia, lo que se relaciona con la capacidad de conmutación máxima del inversor (100 kHz, considerando un paso de muestreo de 10 μs).

Tabla 6.10: Métricas para distintas amplitudes de corriente en el control Deep Q-Network.

A_{ref} [A]	5	15	10	20
THD [%]	2.61	0.89	1.33	0.70
RMSE [%]	1.16	1.48	1.20	1.81

6.2.5. Seguimiento de señal cuadrada

La Figura 6.17 muestra el desempeño del controlador basado en LUT frente a una referencia cuadrada de alta exigencia dinámica. La corriente medida reproduce adecuadamente los niveles superior e inferior de la referencia, sin evidenciar saturación ni desviaciones significativas en régimen estacionario. Durante los flancos de subida y bajada se observa un retardo moderado, atribuible a la dinámica del convertidor y al proceso de conmutación, sin sobreoscilaciones apreciables.

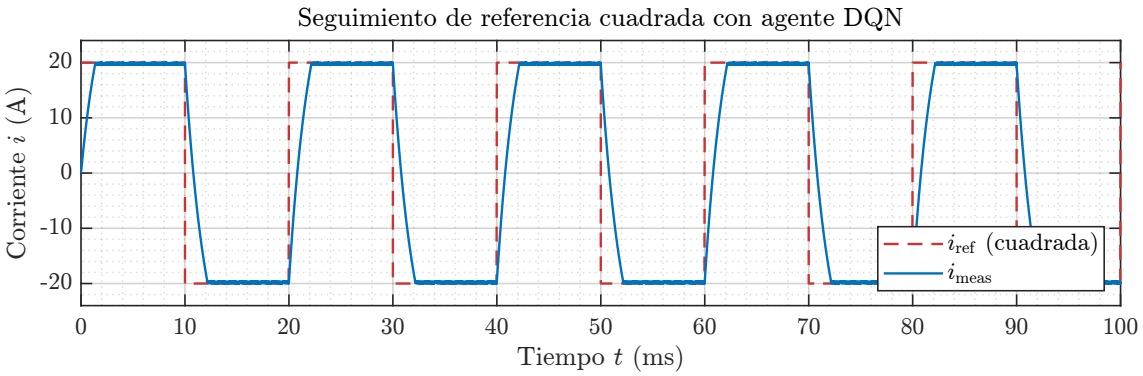


Figura 6.17: Seguimiento de referencia cuadrada - agente DQN en PLECS.

La señal obtenida presenta una forma cuadrada bien definida y una variación mínima alrededor de los niveles deseados, lo cual evidencia que la LUT captura correctamente la política de control incluso ante cambios abruptos en la referencia. Este comportamiento confirma la capacidad del sistema para operar de manera estable y precisa en condiciones de control intensivo.

6.2.6. Impacto de construcción de LUT

La Figura 6.18 ilustra el desempeño del sistema de control al utilizar dos tablas de decisión (*Lookup Tables*) con distintos rangos de corriente. Para el caso de la LUT con límite ± 15 A, se observa saturación cuando la referencia excede dicho umbral, lo que produce un seguimiento truncado y errores sistemáticos en las crestas y valles de la señal. Este efecto se manifiesta como un aplanamiento en los máximos y mínimos, limitando la capacidad del controlador para reproducir la dinámica completa de la corriente solicitada.

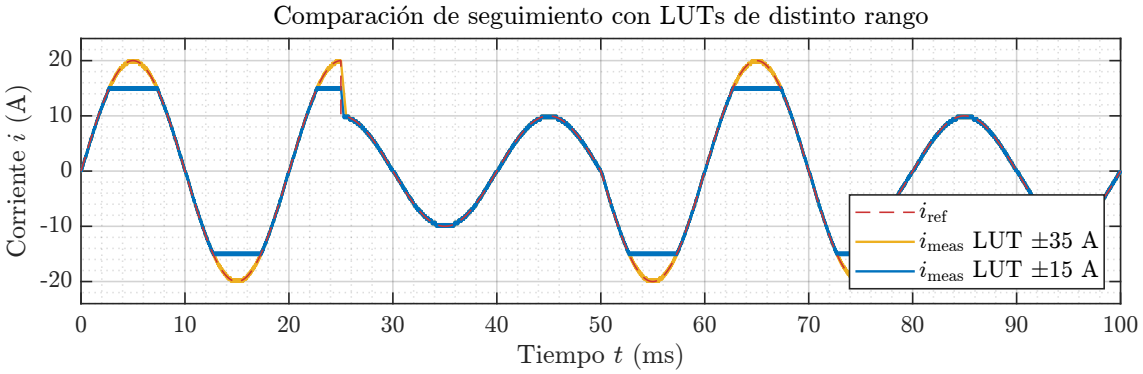


Figura 6.18: Seguimiento de referencia con LUT de diferente rango - agente DQN en PLECS.

Por el contrario, la LUT con rango extendido de ± 35 A permite capturar adecuadamente todo el espectro de amplitudes presentes en la referencia, manteniendo una trayectoria continua y sin saturación. La coincidencia entre referencia y corriente medida es significativamente mejor en los tramos de mayor amplitud, donde el modelo con ± 15 A presenta sus mayores desviaciones.

6.3. Control de corriente trifásico con agente Q-learning marco $\alpha\beta$

La Figura 6.19 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ durante la etapa inicial de operación del agente Q-learning, partiendo desde una tabla Q inicializada en cero y un período de muestreo de $T_s = 10\,\mu s$ ($f_s = 100\,kHz$). En este escenario, el agente no dispone de información previa sobre la política de control, por lo que las acciones aplicadas al inversor se determinan exclusivamente a partir del aprendizaje inmediato, exploración y de la retroalimentación obtenida del sistema eléctrico.

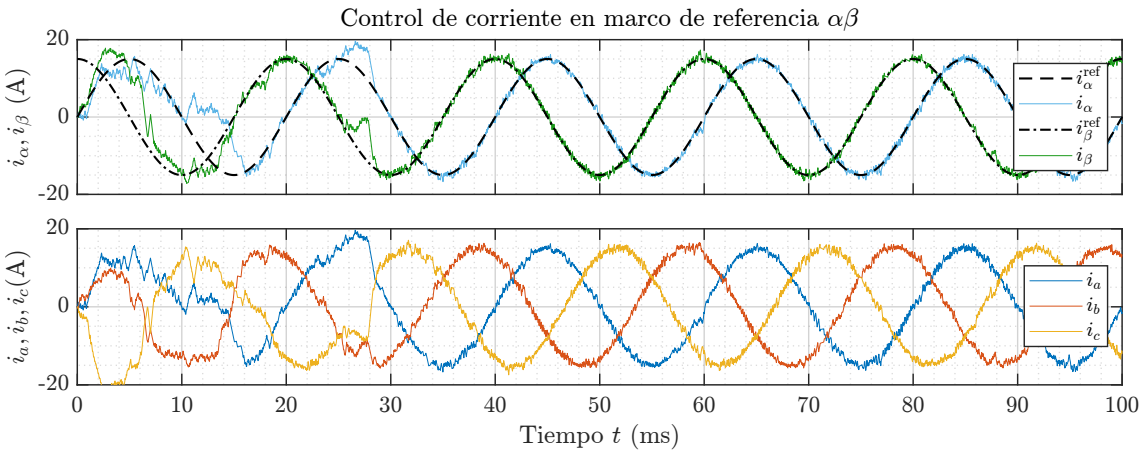


Figura 6.19: Control de corriente en marco de referencia $\alpha\beta$ con agente Q-learning.

En la parte superior de la figura se presentan las corrientes de referencia y las corrientes medidas en el marco $\alpha\beta$, donde se observa que, durante los primeros milisegundos, el seguimiento es impreciso y presenta oscilaciones significativas. Este comportamiento es consistente con la ausencia de conocimiento inicial del agente, que aún no ha aprendido la correspondencia entre el error de corriente y las acciones de conmutación apropiadas.

A medida que transcurre el tiempo, y aproximadamente tras los primeros $30\,\mathrm{ms}$ de operación, el agente comienza a ajustar su política a partir de las recompensas obtenidas, lo que se traduce en una mejora progresiva del seguimiento de las corrientes de referencia. Este proceso de adaptación se refleja en la reducción del desfase y del rizado, así como en la convergencia hacia formas de onda sinusoidales coherentes con la referencia impuesta.

En la parte inferior de la figura se muestran las corrientes trifásicas resultantes, evidenciando que, pese a la condición inicial de aprendizaje desde una tabla Q nula, el control es capaz de estabilizar el sistema y generar un conjunto de corrientes balanceadas una vez que la política comienza a consolidarse. Este resultado confirma la capacidad del agente Q-learning para aprender de manera *online* una estrategia de control efectiva en el marco $\alpha\beta$, sin requerir inicialización previa de la política.

6.3.1. Control de corriente frente a diferentes amplitudes

La Figura 6.20 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ ante variaciones en la amplitud de la corriente de referencia, manteniendo constante la frecuencia de operación. El experimento considera una secuencia de cambios escalonados en la amplitud, comenzando con 20 A, seguido de 10 A, luego 15 A y finalmente 5 A, permitiendo evaluar el comportamiento del agente Q-learning en distintos niveles de exigencia del sistema. El experimento se realiza una vez que la simulación ha transcurrido 100 ms, asegurando que el agente ha superado la etapa inicial de aprendizaje y que la política de control se encuentra en un régimen representativo de operación.

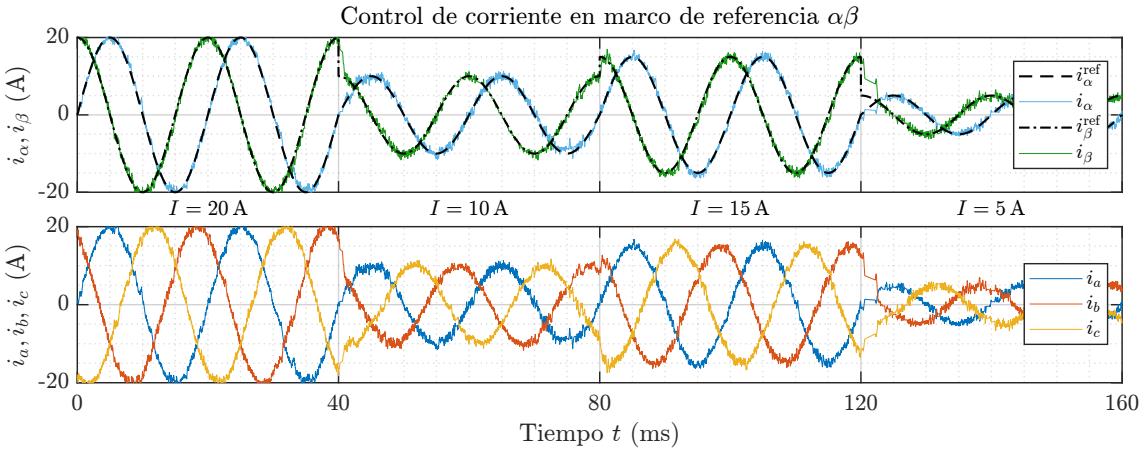


Figura 6.20: Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la amplitud de referencia.

En la parte superior de la figura se presentan las corrientes de referencia y las corrientes medidas en el marco $\alpha\beta$. Se observa que, tras cada cambio de amplitud, el agente requiere un breve intervalo de adaptación, durante el cual aparecen oscilaciones transitorias y un mayor error de seguimiento. No obstante, una vez superada esta etapa, el control logra ajustarse al nuevo nivel de referencia y recuperar un seguimiento coherente con la señal deseada.

La parte inferior de la figura muestra las corrientes trifásicas correspondientes, donde se aprecia que el sistema mantiene un comportamiento balanceado en las tres fases a lo largo de toda la prueba. Para amplitudes mayores, el seguimiento presenta una mejor definición de las formas de onda, mientras que al reducir la amplitud a 5 A se observa un aumento relativo del rizado y de la dispersión de la corriente, atribuible a la discretización del espacio de estados y a la menor relación señal–ruido en rangos de corriente reducidos.

Estos resultados indican que el agente Q-learning es capaz de adaptarse a cambios significativos en la amplitud de la referencia y mantener la estabilidad del control en el marco $\alpha\beta$. Sin embargo, el desempeño relativo se ve afectado en amplitudes más bajas, lo que evidencia una mayor sensibilidad del método tabular a errores de cuantización cuando el sistema opera cerca del origen del espacio de estados.

6.3.2. Control de corriente frente a variaciones de resistencia de carga

La Figura 6.21 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ frente a variaciones abruptas en el valor de la resistencia de carga, manteniendo constante la inductancia y la frecuencia de la referencia. El experimento se realiza una vez superada la etapa inicial de aprendizaje del agente, de modo que la política de control se encuentra en un régimen representativo de operación.

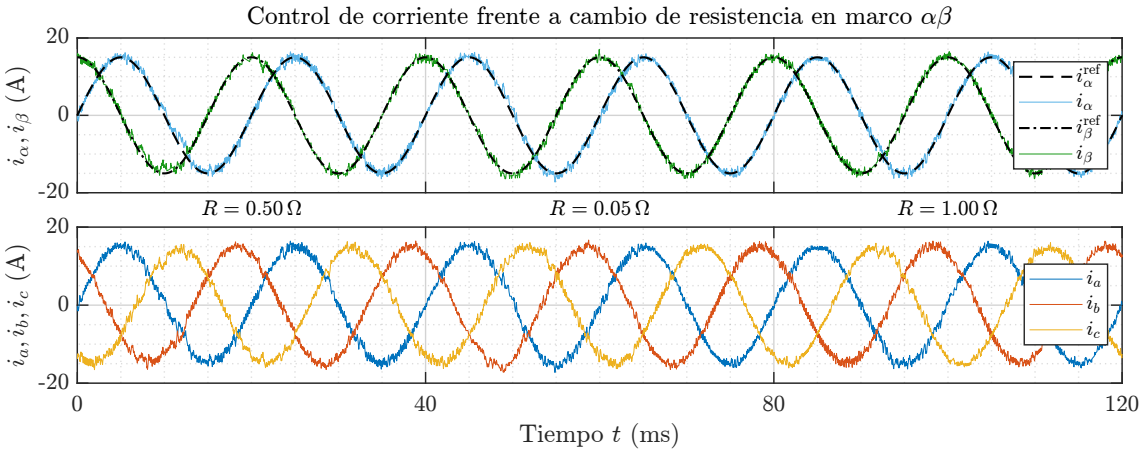


Figura 6.21: Respuesta del control de corriente en marco de referencia $\alpha\beta$ ante cambios en la resistencia.

Durante la prueba, la resistencia de carga se modifica de manera escalonada desde R = 0.50 Ω hasta R = 0.05 Ω, lo que corresponde a una disminución del 90 % respecto de su valor inicial. Posteriormente, la resistencia se incrementa hasta R = 1.00 Ω, equivalente a un aumento del 100 % en relación con el valor original. Estas variaciones permiten evaluar el comportamiento del agente frente a cambios significativos en la dinámica resistiva del sistema.

En la parte superior de la figura se presentan las corrientes de referencia y las corrientes medidas en el marco $\alpha\beta$. Se observa que, a pesar de las variaciones abruptas en el valor de la resistencia, el seguimiento de la corriente se mantiene prácticamente inalterado, sin evidenciar cambios relevantes en la amplitud ni en la fase de la señal medida. No se aprecian transitorios pronunciados ni degradación sostenida del seguimiento tras los cambios de carga.

La parte inferior de la figura muestra las corrientes trifásicas correspondientes, donde se aprecia que el sistema conserva un comportamiento balanceado en las tres fases a lo largo de toda la prueba. La ausencia de un impacto notorio ante variaciones resistivas de gran magnitud sugiere que el agente Q-learning logra compensar de manera efectiva los cambios en la componente resistiva de la carga, manteniendo una respuesta estable del sistema en el marco de referencia $\alpha\beta$.

6.3.3. Control de corriente frente a variaciones de inductancia de carga

La Figura 6.22 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ frente a variaciones abruptas en el valor de la inductancia de carga, manteniendo constante la resistencia y la frecuencia de la corriente de referencia. Al igual que en los experimentos anteriores, las pruebas se realizan una vez superada la etapa inicial de aprendizaje del agente, de modo que la política de control se encuentra en un régimen representativo de operación.

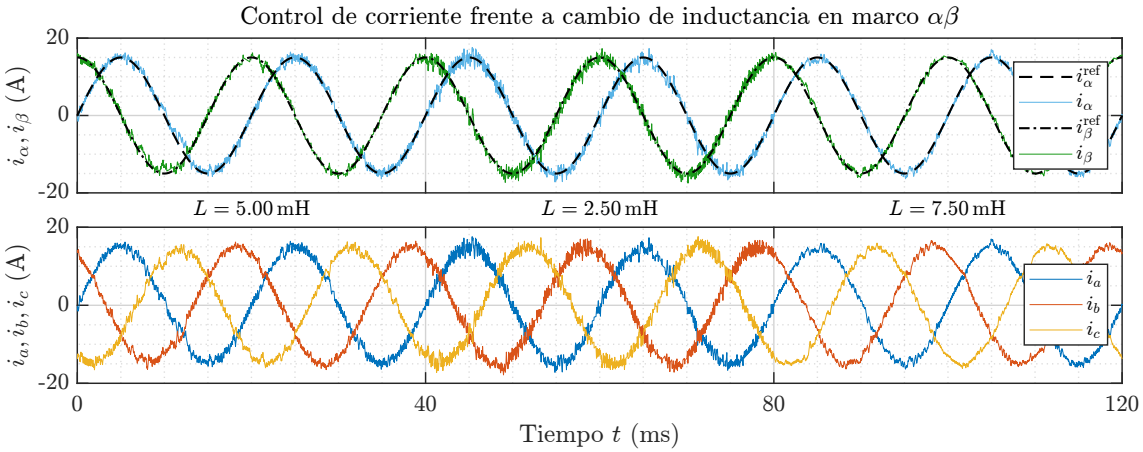


Figura 6.22: Respuesta del control de corriente en marco de referencia $\alpha\beta$ ante cambios en la inductancia.

Durante el experimento, la inductancia de la carga se modifica de manera escalonada desde L = 5.00 mH hasta L = 2.50 mH, lo que corresponde a una disminución del $50\,\%$ respecto de su valor inicial. Posteriormente, la inductancia se incrementa hasta L = 7.50 mH, equivalente a un aumento del $50\,\%$ en relación con el valor original. Estas variaciones permiten analizar el efecto de la dinámica inductiva sobre el desempeño del control.

En la parte superior de la figura se presentan las corrientes de referencia y las corrientes medidas en el marco $\alpha\beta$. Se observa que, al reducir la inductancia, el seguimiento de la corriente presenta un aumento visible del rizado, asociado a una mayor rapidez natural del sistema y a un incremento en la pendiente de la corriente frente a las acciones de conmutación del inversor. Este efecto es coherente con la reducción de la constante de tiempo eléctrica del sistema RL.

Al incrementar la inductancia hasta 7.50 mH, el rizado disminuye nuevamente y las formas de onda recuperan una mayor suavidad, evidenciando una mejor atenuación de las componentes de alta frecuencia introducidas por la conmutación. En la parte inferior de la figura, las corrientes trifásicas se mantienen balanceadas durante toda la prueba, aunque con un contenido de rizado mayor en el caso de menor inductancia.

Estos resultados muestran que, a diferencia de las variaciones resistivas, los cambios en la inductancia de carga tienen un impacto directo y apreciable sobre la calidad de la forma de onda de la corriente. No obstante, el agente Q-learning es capaz de mantener la estabilidad del control y el seguimiento global de la referencia en todo el rango de inductancias considerado, adaptándose a la modificación de la dinámica del sistema en el marco $\alpha\beta$.

6.3.4. Control de corriente frente a variaciones de frecuencia eléctrica

La Figura 6.23 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ frente a incrementos escalonados en la frecuencia eléctrica de la señal de referencia. El experimento se realiza una vez que el agente Q-learning ha superado la etapa inicial de aprendizaje, de modo que la política de control se encuentra estabilizada antes de aplicar las variaciones de frecuencia.

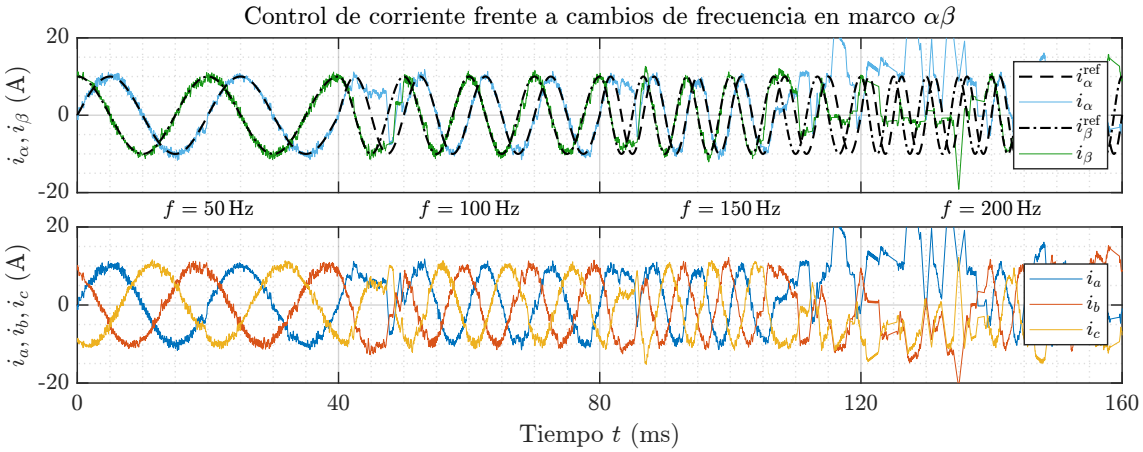


Figura 6.23: Desempeño del control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la frecuencia eléctrica.

Durante la prueba, la frecuencia de la corriente de referencia se incrementa progresivamente desde 50 Hz hasta 100 Hz, 150 Hz y finalmente 200 Hz, manteniendo constante la amplitud de la señal. Esta secuencia permite evaluar la capacidad del agente para adaptarse a dinámicas cada vez más rápidas, sin modificar los parámetros eléctricos de la carga.

En la parte superior de la figura se observan las corrientes de referencia y las corrientes medidas en el marco $\alpha\beta$. Para 50 Hz y 100 Hz, el control logra un seguimiento adecuado, con formas de onda bien definidas y un rizado acotado. Al aumentar la frecuencia a 150 Hz, comienzan a apreciarse oscilaciones adicionales y una mayor dispersión en la corriente medida, reflejando el aumento de la exigencia dinámica impuesta al inversor.

A 200 Hz, el desempeño del control se degrada de manera significativa. En este régimen, el agente Q-learning pierde estabilidad y no logra reproducir de forma consistente la referencia, observándose oscilaciones marcadas y comportamientos erráticos tanto en el marco $\alpha\beta$ como en las corrientes trifásicas. Este fenómeno no se explica únicamente por las limitaciones físicas del inversor, sino que está directamente asociado a la incapacidad del algoritmo para adaptarse a una dinámica que evoluciona demasiado rápido respecto de la frecuencia de muestreo y del proceso de aprendizaje discreto.

En particular, a frecuencias elevadas, la información disponible para el agente en cada paso temporal resulta insuficiente para capturar adecuadamente la evolución del sistema, lo que conduce a actualizaciones de la tabla Q poco representativas y a una política inestable.

6.4. Control de corriente trifásico con agente Q-learning marco *dq*

En esta sección se aborda el control de corriente trifásico mediante un agente Q-learning formulado en el marco de referencia síncrono dq. A diferencia del caso en marco $\alpha\beta$, la implementación en coordenadas dq presenta mayores exigencias desde el punto de vista del aprendizaje y de la discretización del espacio de estados, debido a la dinámica más rápida de las variables transformadas y a la fuerte dependencia temporal introducida por la rotación del sistema de referencia.

Durante el desarrollo de este estudio se constató que no fue posible obtener un comportamiento estable del control utilizando un período de muestreo del orden de $T_s = 10\,\mu$s (100 kHz), aun considerando distintas combinaciones de hiperparámetros y mecanismos de estabilización de la política. En todos estos casos, el agente mostró dificultades para adaptarse de manera consistente a la dinámica del sistema, lo que se tradujo en oscilaciones persistentes o pérdida de seguimiento de la referencia. Para lograr un desempeño aceptable fue necesario reducir el período de muestreo a $T_s = 1\,\mu$s (1 MHz) y, adicionalmente, disminuir la dimensión de la tabla Q mediante una discretización más acotada del espacio de estados. Estas modificaciones permitieron que el agente capturara con mayor fidelidad la evolución temporal de las corrientes en el marco dq, posibilitando el aprendizaje de una política funcional aun cuando la tabla Q se inicializa en cero.

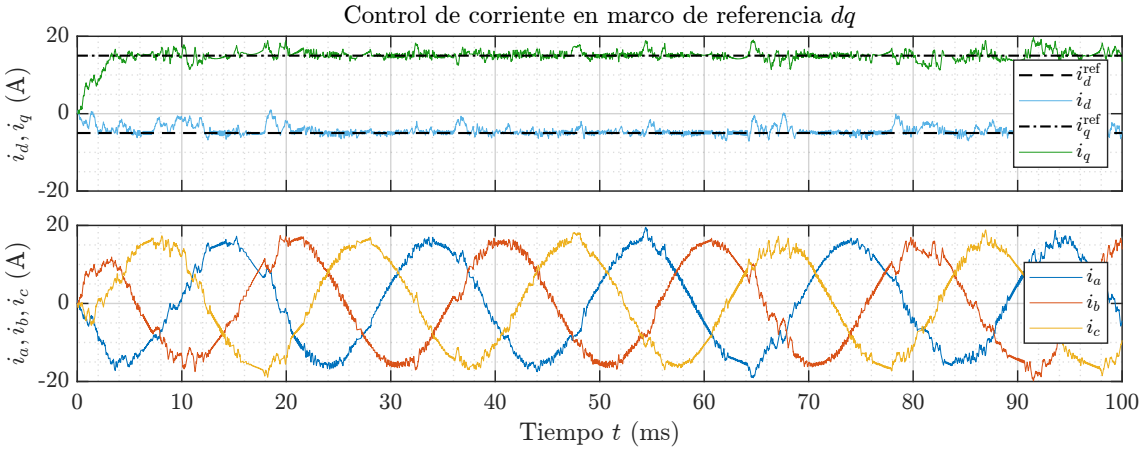


Figura 6.24: Control de corriente en marco de referencia dq con agente Q-learning.

La Figura 6.24 muestra la respuesta del control de corriente trifásico en el marco de referencia dq durante la etapa inicial de operación del agente Q-learning, partiendo desde una tabla Q inicializada en cero. En la parte superior se presentan las componentes i_d e i_q, junto con sus respectivas referencias, mientras que en la parte inferior se muestran las corrientes trifásicas resultantes en el sistema.

Durante los primeros milisegundos se observa un régimen transitorio marcado por fluctuaciones en las variables dq, particularmente en la componente i_q, lo cual es consistente con la ausencia de información previa en la política del agente. A medida que avanza el aprendizaje, las corrientes en el marco dq convergen progresivamente hacia sus valores de referencia, estabilizándose alrededor de los 20–30 ms.

6.4.1. Control de corriente frente a diferentes amplitudes

La Figura 6.25 muestra la respuesta del control de corriente trifásico en el marco de referencia dq frente a variaciones escalonadas en la amplitud de la corriente de referencia. El experimento se realiza una vez que el agente ha superado la etapa inicial de aprendizaje, permitiendo evaluar su capacidad de adaptación ante cambios en el nivel de operación bajo una política ya parcialmente consolidada.

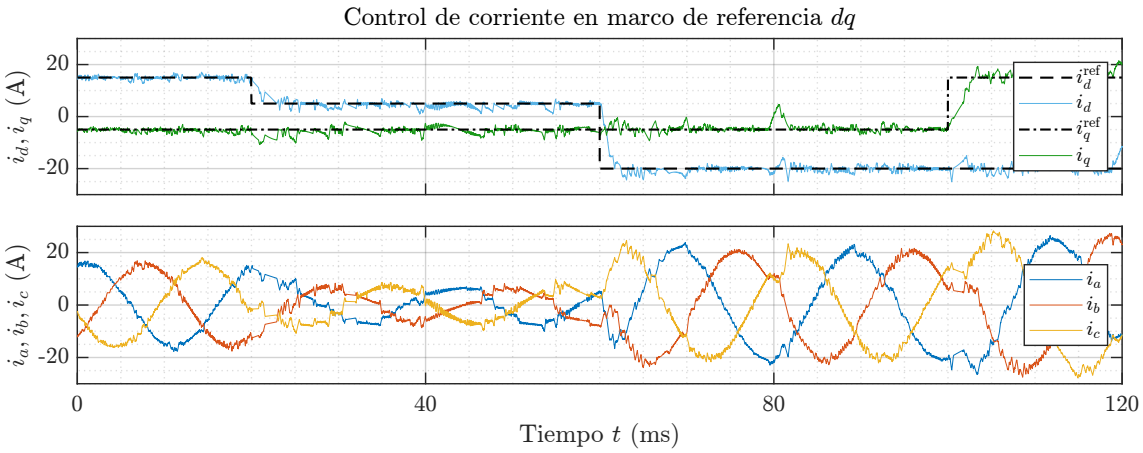


Figura 6.25: Control de corriente con agente Q-learning en marco de referencia dq.

En la parte superior de la figura se presentan las componentes i_d e i_q junto con sus referencias, donde se observa que el agente logra ajustar la corriente al nuevo nivel impuesto tras cada cambio de amplitud. Sin embargo, durante los transitorios asociados a dichas variaciones se evidencian desviaciones apreciables y un aumento del rizado, particularmente en la componente i_q, lo que refleja la dificultad del agente para reaccionar de forma inmediata ante cambios bruscos en el marco síncrono.

En la parte inferior se muestran las corrientes trifásicas resultantes, donde se aprecia que, si bien el sistema mantiene corrientes balanceadas una vez alcanzado el nuevo régimen estacionario, la calidad de la forma de onda se ve afectada durante los intervalos de transición. Estas perturbaciones son más notorias en los cambios hacia amplitudes mayores, donde la exigencia dinámica sobre el controlador se incrementa.

Estos resultados muestran que el agente Q-learning en marco dq es capaz de adaptarse a distintas amplitudes de referencia una vez superada la etapa inicial de aprendizaje. No obstante, la respuesta transitoria presenta un mayor nivel de rizado y desviación respecto a la referencia en comparación con el caso en $\alpha\beta$, especialmente durante cambios abruptos de amplitud.

Este comportamiento pone de manifiesto las limitaciones del enfoque tabular en el marco síncrono, donde la combinación entre discretización del espacio de estados y dinámica rápida del sistema obliga a utilizar períodos de muestreo muy reducidos para preservar la estabilidad del control.

6.4.2. Control de corriente frente a variaciones de resistencia de carga

La Figura 6.26 muestra la respuesta del control de corriente trifásico en el marco de referencia dq ante variaciones escalonadas en la resistencia de la carga. El experimento se realiza una vez que el agente ha alcanzado un régimen de operación estable, permitiendo evaluar el efecto de cambios en la dinámica resistiva sobre una política de control ya aprendida.

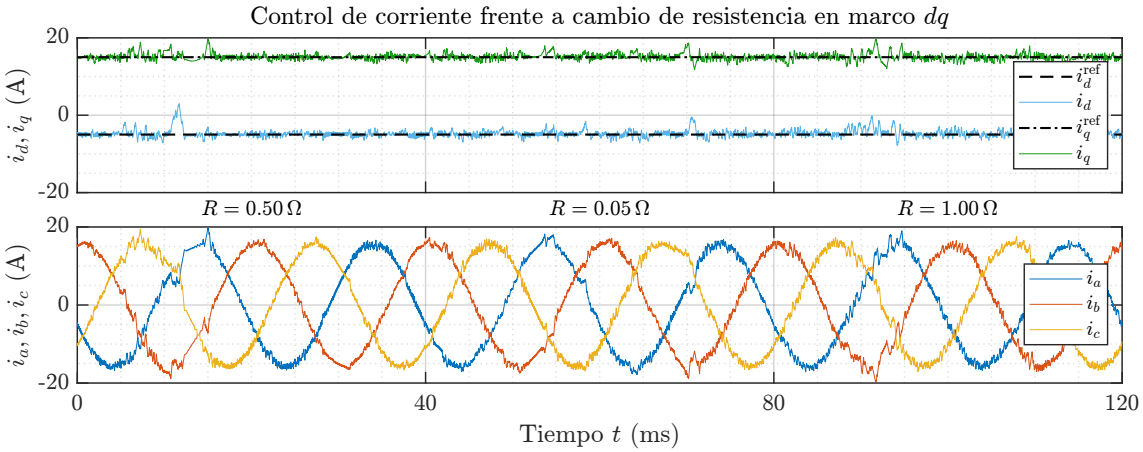


Figura 6.26: Control de corriente con agente Q-learning en marco de referencia dq.

En la parte superior de la figura se presentan las componentes i_d e i_q junto con sus respectivas referencias. Se observa que los cambios en la resistencia, que corresponden a variaciones relativas significativas respecto al valor nominal, no generan desviaciones apreciables ni pérdida de seguimiento en el marco dq. Las componentes de corriente permanecen próximas a sus valores de referencia durante todo el experimento.

Las corrientes trifásicas mostradas en la parte inferior mantienen su carácter balanceado y sinusoidal para todas las configuraciones de resistencia consideradas. A diferencia de lo observado en los cambios de amplitud, las transiciones entre distintos valores de resistencia no introducen perturbaciones relevantes en la forma de onda ni incrementos notorios del rizado.

Este comportamiento es consistente con el caso en marco $\alpha\beta$ y se explica por el hecho de que las variaciones puramente resistivas modifican la magnitud de la impedancia sin alterar de forma significativa la constante de tiempo del sistema, permitiendo que el agente mantenga la política de control aprendida sin requerir una adaptación sustantiva.

6.4.3. Control de corriente frente a variaciones de inductancia de carga

La Figura 6.27 muestra la respuesta del control de corriente trifásico en el marco de referencia dq ante variaciones escalonadas en la inductancia de la carga. Al igual que en el caso formulado en $\alpha\beta$, este experimento permite evaluar el efecto de cambios en la constante de tiempo del sistema sobre una política de control previamente aprendida.

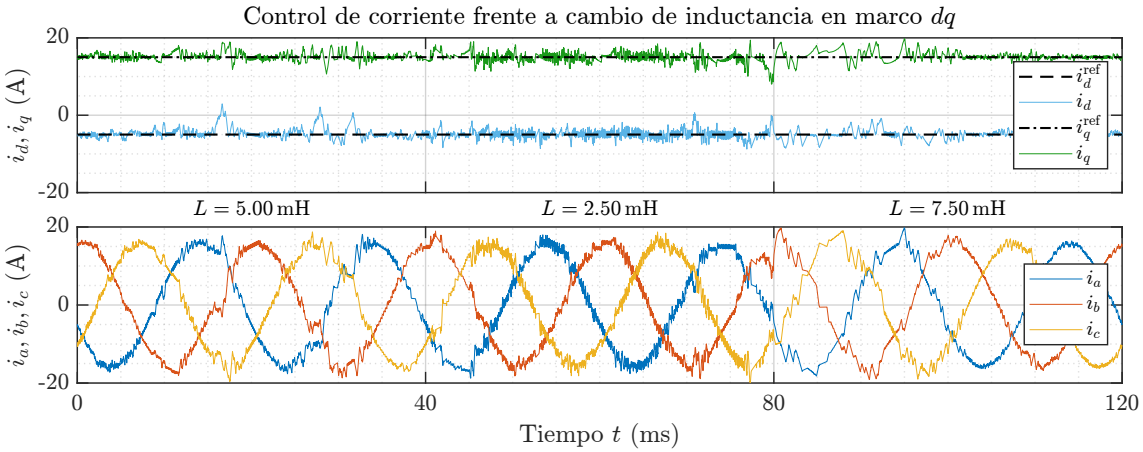


Figura 6.27: Control de corriente en marco de referencia dq con agente Q-learning.

En la parte superior de la figura se presentan las componentes i_d e i_q junto con sus referencias. Al reducir la inductancia desde su valor nominal, se observa un aumento del rizado en ambas componentes, particularmente en i_q, reflejando la mayor rapidez de la dinámica eléctrica. Al incrementar la inductancia, la respuesta se vuelve más suave y con menor variación instantánea, evidenciando el efecto amortiguador de una constante de tiempo mayor.

Las corrientes trifásicas mostradas en la parte inferior exhiben un comportamiento consistente con estas observaciones. Para valores bajos de inductancia, el rizado aumenta de forma apreciable y la forma de onda se vuelve menos suave, mientras que al aumentar la inductancia las corrientes recuperan una apariencia más regular y cercana a una sinusoide ideal.

Este comportamiento reproduce lo observado en el marco $\alpha\beta$, confirmando que las variaciones inductivas tienen un impacto directo sobre la calidad de la corriente controlada. No obstante, en el marco dq estas perturbaciones resultan más visibles debido a la sensibilidad del agente Q-learning a dinámicas rápidas, especialmente considerando la discretización del espacio de estados y la elevada frecuencia de muestreo requerida para mantener la estabilidad del control.

6.4.4. Control de corriente frente a variaciones de frecuencia eléctrica

La Figura 6.28 presenta la respuesta del control de corriente trifásico en el marco de referencia dq frente a incrementos escalonados en la frecuencia eléctrica de la referencia. El experimento se inicia en régimen estacionario a 50 Hz y posteriormente se incrementa la frecuencia a 100 Hz y 150 Hz, manteniendo constante la amplitud de la corriente solicitada.

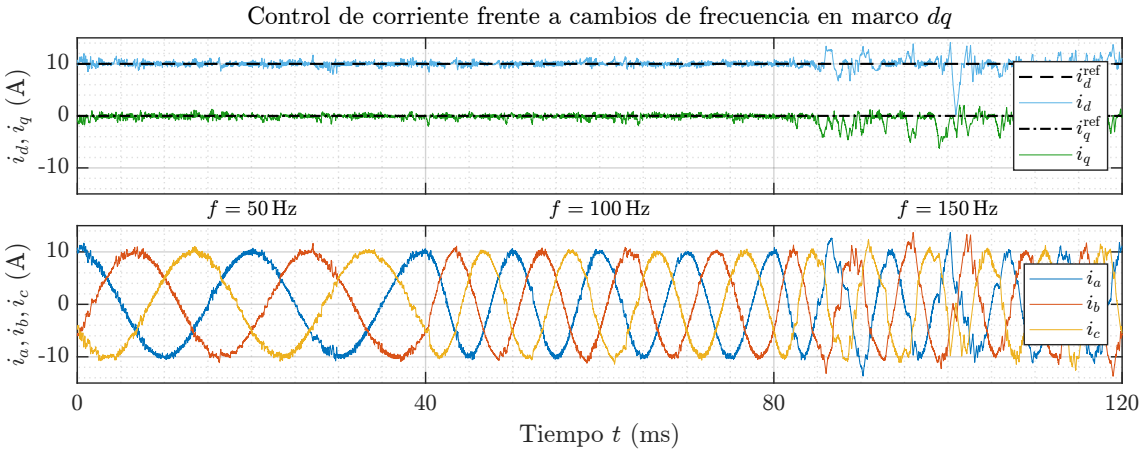


Figura 6.28: Control de corriente en marco de referencia dq con agente Q-learning.

En la parte superior de la figura se muestran las componentes i_d e i_q junto con sus referencias. Para frecuencias bajas, el agente logra mantener las corrientes próximas a los valores deseados, aunque con un aumento progresivo del rizado a medida que la frecuencia se incrementa. Al alcanzar los 150 Hz, se observa una degradación significativa del seguimiento, con oscilaciones abruptas y pérdida de estabilidad en ambas componentes.

Este comportamiento se refleja directamente en las corrientes trifásicas mostradas en la parte inferior de la figura. Mientras que a 50 Hz y 100 Hz las corrientes conservan una forma aproximadamente sinusoidal y balanceada, al aumentar la frecuencia aparecen distorsiones severas, desbalance y variaciones de amplitud que evidencian la incapacidad del agente para sostener el control en este régimen.

La pérdida de estabilidad observada no se asocia a una saturación del inversor ni a un límite físico del sistema, sino a la dificultad del agente Q-learning para adaptarse a la dinámica acelerada impuesta por el marco dq. En particular, el aprendizaje tabular resulta insuficiente para capturar de manera consistente la evolución del sistema bajo variaciones rápidas de frecuencia, aun cuando se emplea un período de muestreo extremadamente reducido.

6.5. Control de corriente trifásico con agente DQN en marco $\alpha\beta$

En esta sección se analiza el desempeño del control de corriente trifásico en el marco de referencia $\alpha\beta$ utilizando un agente *Deep Q-Network*. El controlador ejecuta directamente la política aprendida durante la etapa de entrenamiento en MATLAB, la cual es implementada en *PLECS* para su evaluación en régimen. A diferencia del enfoque basado en Q-learning tabular, el agente DQN no realiza aprendizaje *online*, lo que permite evaluar su comportamiento desde el inicio de la simulación sin fases transitorias asociadas a exploración o adaptación.

La Figura 6.29 muestra la respuesta inicial del sistema para un período de muestreo de $T_s = 10\,\mu$s. En la parte superior se presentan las corrientes de referencia y las corrientes medidas en el marco $\alpha\beta$, donde se observa un seguimiento estable y bien definido desde los primeros instantes. Las componentes α y β reproducen correctamente las formas de onda sinusoidales esperadas, manteniendo coherencia tanto en amplitud como en desfase respecto de sus referencias.

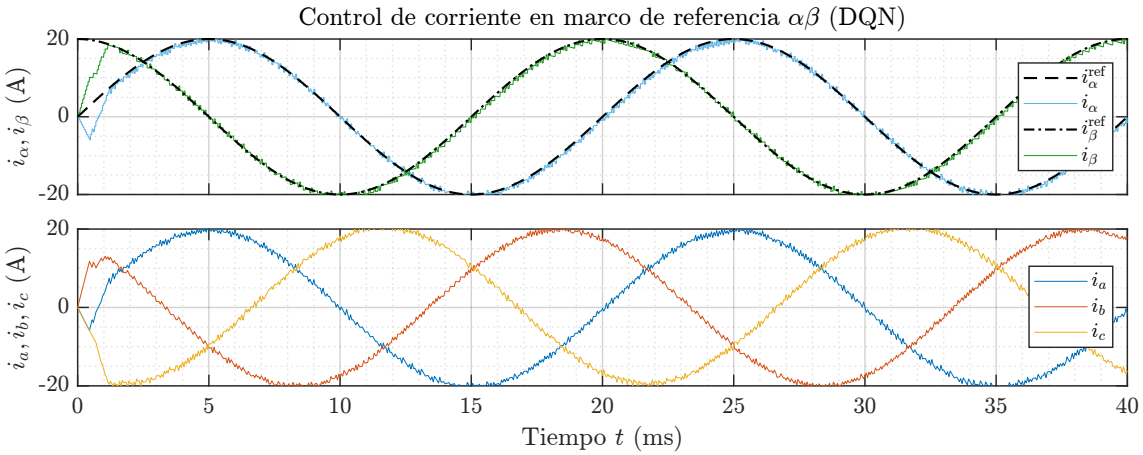


Figura 6.29: Control de corriente en marco de referencia $\alpha\beta$.

En la parte inferior de la figura se muestran las corrientes trifásicas resultantes, las cuales presentan un comportamiento balanceado y suave desde el inicio del control. No se evidencian oscilaciones transitorias ni inestabilidades iniciales, lo que indica que la política implementada captura adecuadamente la dinámica del sistema trifásico en el marco $\alpha\beta$ incluso a frecuencias de muestreo elevadas.

6.5.1. Control de corriente frente a diferentes amplitudes

La Figura 6.30 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ ante variaciones escalonadas en la amplitud de la corriente de referencia. El experimento se realiza una vez alcanzado el régimen estacionario y considera cambios sucesivos de amplitud entre 20 A, 10 A, 15 A y 5 A, manteniendo constante la frecuencia eléctrica.

En la parte superior de la figura se observa que las componentes α y β siguen adecuadamente las referencias impuestas en todos los niveles de amplitud. Tras cada cambio, el controlador ajusta la magnitud de la corriente sin introducir oscilaciones adicionales ni transitorios prolongados, manteniendo una forma de onda sinusoidal bien definida incluso en los tramos de menor amplitud.

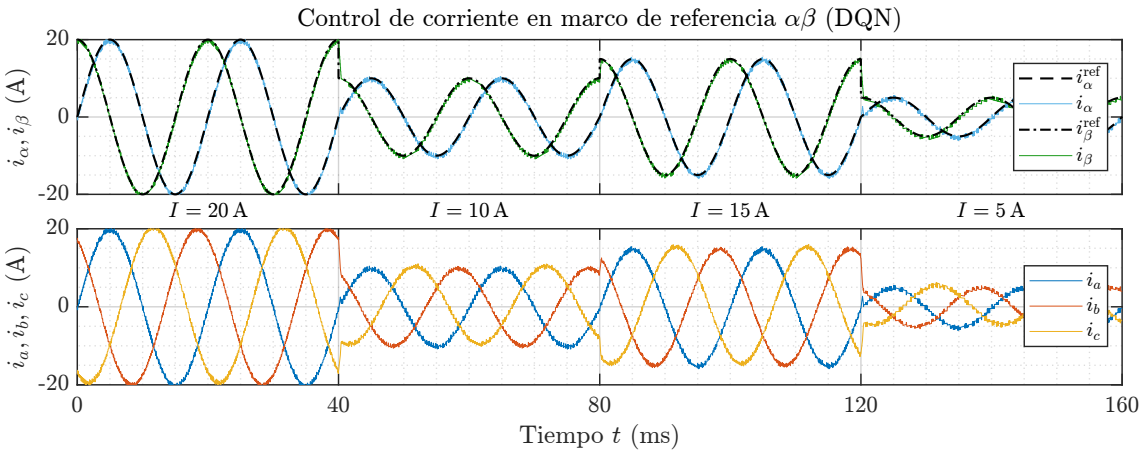


Figura 6.30: Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la amplitud de referencia.

Las corrientes trifásicas mostradas en la parte inferior conservan un comportamiento balanceado durante todo el experimento. A medida que disminuye la amplitud de referencia, se aprecia un leve aumento del rizado relativo, asociado a la resolución finita de la política discretizada, sin comprometer la estabilidad del control ni la simetría entre fases.

6.5.2. Control de corriente frente a variaciones de resistencia de carga

La Figura 6.31 presenta la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ ante variaciones abruptas en la resistencia de la carga. El experimento se realiza en régimen estacionario y considera cambios sucesivos del valor resistivo desde 0.50 Ω a 0.05 Ω y posteriormente a 1.00 Ω, manteniendo constante la amplitud y la frecuencia de la corriente de referencia.

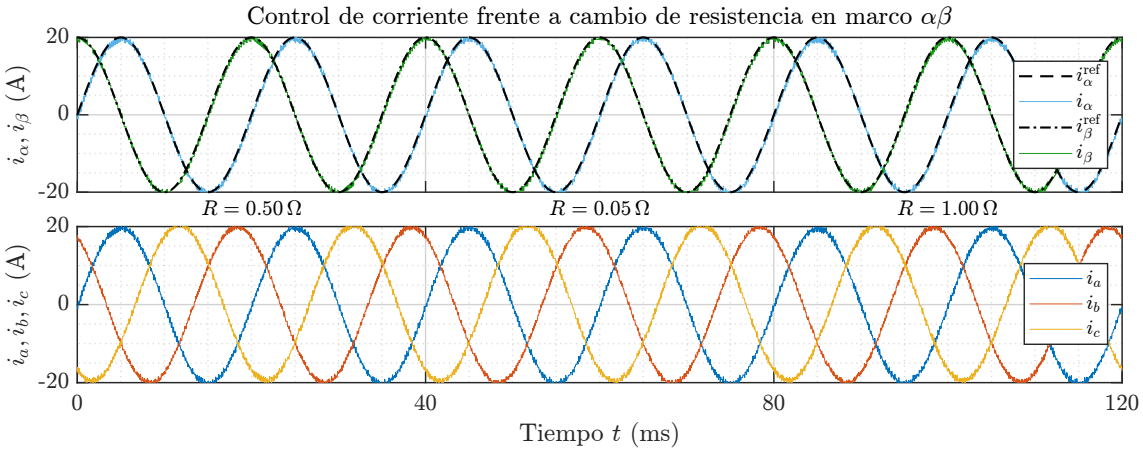


Figura 6.31: Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la resistencia de la carga.

En la parte superior de la figura se observa que las componentes α y β mantienen un seguimiento preciso de sus respectivas referencias durante todo el experimento. Los cambios en la resistencia no introducen desviaciones apreciables en amplitud ni desfases relevantes, lo que evidencia una baja sensibilidad del control frente a variaciones resistivas significativas.

Las corrientes trifásicas mantienen una forma sinusoidal balanceada en los tres tramos evaluados. A pesar de que la variación desde 0.05 Ω a 1.00 Ω representa un cambio de un orden de magnitud

en el valor resistivo, no se observan efectos notorios en el rizado ni en la estabilidad del sistema.

6.5.3. Control de corriente frente a variaciones de inductancia de carga

La Figura 6.32 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ ante variaciones escalonadas en la inductancia de la carga, considerando valores de 5.00 mH, 2.50 mH y 7.50 mH.

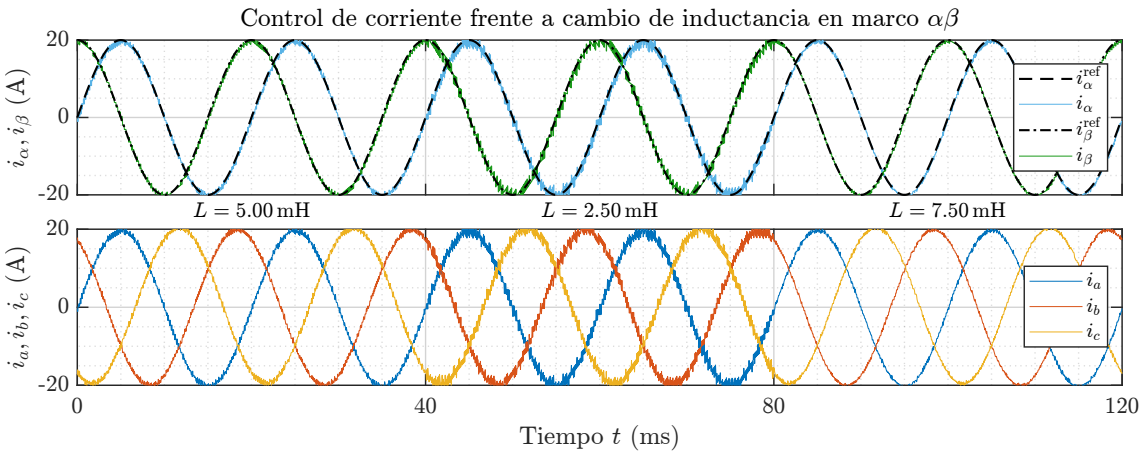


Figura 6.32: Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la inductancia de la carga.

Al reducir la inductancia se observa un aumento del rizado en las corrientes medidas, asociado a la mayor rapidez de la dinámica del sistema y a la disminución del efecto de filtrado natural de la carga. Al incrementar la inductancia, el rizado se atenúa nuevamente, reflejando una respuesta más suavizada.

Las corrientes trifásicas mantienen un comportamiento balanceado y estable en los tres escenarios evaluados. Si bien la reducción de la inductancia incrementa el contenido de alta frecuencia, el controlador conserva el seguimiento de la amplitud impuesta sin pérdida de sincronismo entre fases.

6.5.4. Control de corriente frente a variaciones de frecuencia eléctrica

La Figura 6.33 muestra la respuesta del control de corriente trifásico en el marco de referencia $\alpha\beta$ frente a incrementos escalonados en la frecuencia eléctrica de la referencia, desde 50 Hz hasta 300 Hz, manteniendo constante la amplitud de la corriente.

Las componentes α y β mantienen un seguimiento coherente de sus referencias a lo largo de todo el barrido de frecuencia. A medida que la frecuencia aumenta, las corrientes presentan una mayor densidad de oscilaciones, consistente con la reducción del período de la señal y con las limitaciones dinámicas propias del convertidor.

Las corrientes trifásicas conservan un comportamiento balanceado durante todo el experimento, sin evidenciar pérdida de estabilidad ni distorsiones abruptas incluso a las frecuencias más elevadas consideradas.

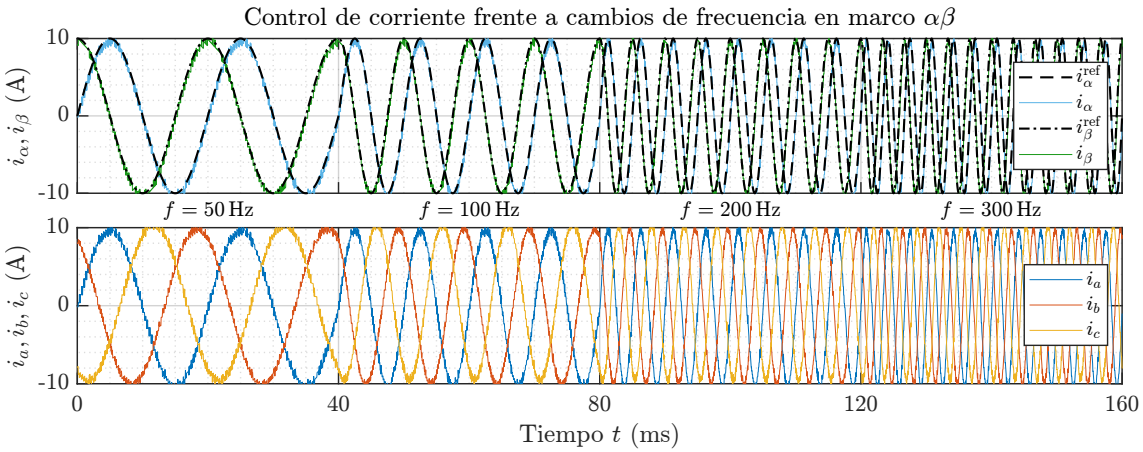


Figura 6.33: Control de corriente en marco de referencia $\alpha\beta$ frente a cambios en la frecuencia de referencia.

6.6. Control de corriente trifásico con agente DQN marco *dq*

En esta sección se analiza el desempeño del control de corriente trifásico en el marco de referencia síncrono dq utilizando un agente *Deep Q-Network*. A diferencia de lo observado en el caso del control basado en Q-learning en este mismo marco, no es necesario reducir el período de muestreo para lograr un funcionamiento estable. El controlador DQN permite operar con un período de muestreo de $T_s = 10\,\mu$s, manteniendo un seguimiento adecuado de las corrientes y una dinámica consistente desde el inicio de la simulación.

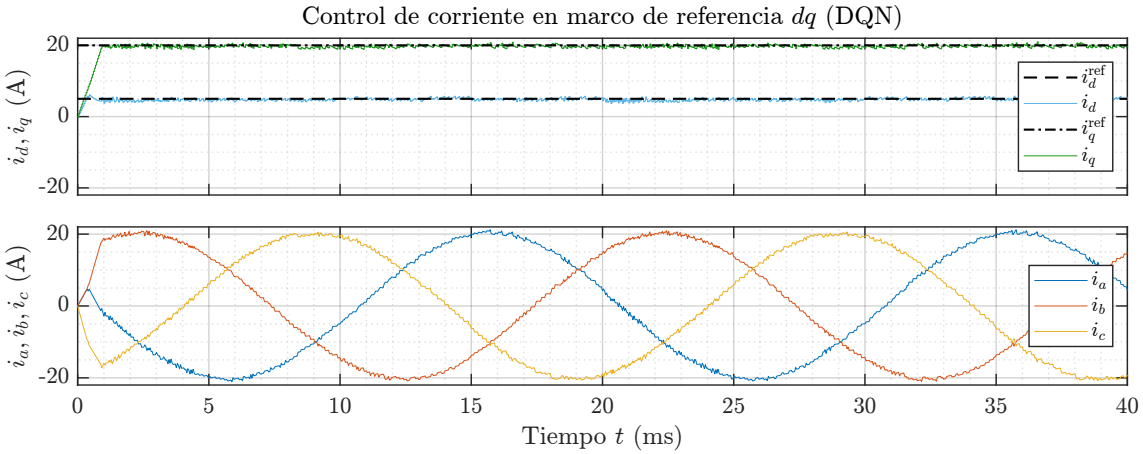


Figura 6.34: Control de corriente en marco de referencia dq.

La Figura 6.34 muestra la respuesta inicial del control de corriente en el marco de referencia dq. En la parte superior se presentan las componentes i_d e i_q junto a sus respectivas referencias. Se observa que ambas corrientes convergen rápidamente a los valores deseados, sin oscilaciones apreciables ni transitorios prolongados, lo que indica una correcta regulación en el marco síncrono.

En la parte inferior de la figura se muestran las corrientes trifásicas resultantes, las cuales presentan formas de onda sinusoidales balanceadas desde los primeros instantes de operación. La ausencia de perturbaciones visibles durante el arranque refleja que la política aprendida por el agente DQN captura de forma adecuada la dinámica del sistema en el marco dq, permitiendo una ejecución estable sin requerir ajustes adicionales en el tiempo de muestreo.

6.6.1. Control de corriente frente a diferentes amplitudes

La Figura 6.35 muestra la respuesta del control de corriente en el marco de referencia dq ante variaciones escalonadas en la amplitud de la corriente de referencia. El experimento considera cambios sucesivos del valor pico de la referencia, manteniendo constante la frecuencia eléctrica.

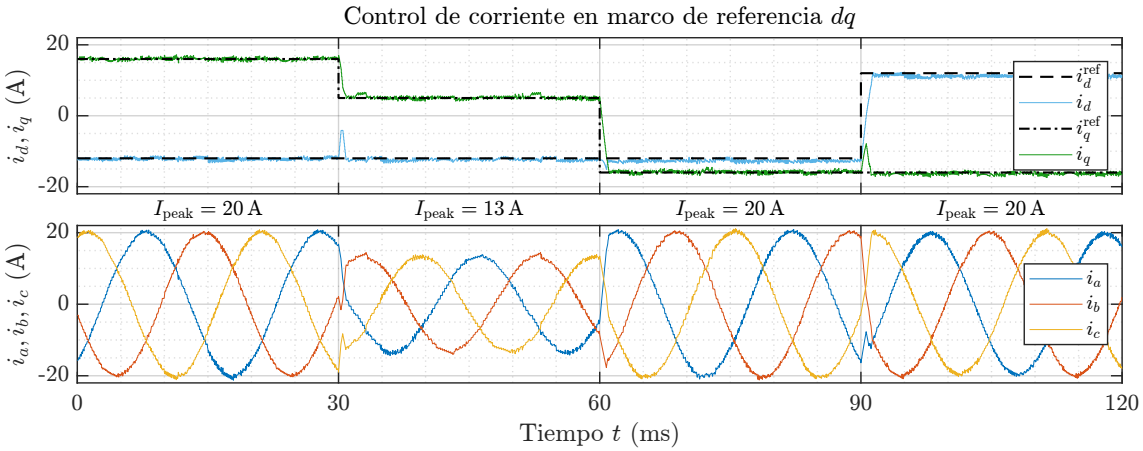


Figura 6.35: Control de corriente en marco de referencia dq frente a cambios en la amplitud de referencia.

En la parte superior de la figura se observa que la componente asociada a la referencia activa se ajusta rápidamente al nuevo nivel impuesto en cada escalón, alcanzando el valor deseado sin oscilaciones prolongadas. Sin embargo, durante los instantes de cambio de amplitud se aprecia una perturbación transitoria en la componente ortogonal, la cual se manifiesta como un desvío breve respecto de su valor de referencia.

Este acoplamiento transitorio entre las componentes d y q es particularmente visible en los cambios abruptos de amplitud, donde la variación de una componente introduce una perturbación momentánea en la otra. No obstante, dicha interacción se atenúa rápidamente, y ambas corrientes convergen nuevamente a sus valores de referencia, lo que indica que el controlador logra restablecer el desacople propio del marco síncrono tras el transitorio.

En la parte inferior de la figura se muestran las corrientes trifásicas resultantes, las cuales mantienen formas de onda sinusoidales balanceadas a lo largo de todo el experimento. Los cambios de amplitud se reflejan de manera coherente en las tres fases, sin introducir distorsiones persistentes ni pérdida de simetría, lo que confirma que las perturbaciones observadas en el marco dq se limitan a transitorios breves y no comprometen el comportamiento global del sistema.

6.6.2. Control de corriente frente a variaciones de resistencia de carga

La Figura 6.36 muestra la respuesta del control de corriente en el marco de referencia dq frente a variaciones escalonadas en la resistencia de la carga. El experimento se realiza en régimen estacionario y considera cambios sucesivos del valor resistivo desde 0.50 Ω a 0.05 Ω y posteriormente a 1.00 Ω, manteniendo constante la amplitud y la frecuencia de la corriente de referencia.

En la parte superior de la figura se observa que las componentes i_d e i_q permanecen prácticamente inalteradas durante todo el experimento. Los cambios en la resistencia no generan perturbaciones apreciables ni transitorios relevantes en ninguna de las componentes, lo que indica que el controlador mantiene una regulación efectiva aun frente a grandes variaciones de resistencia.

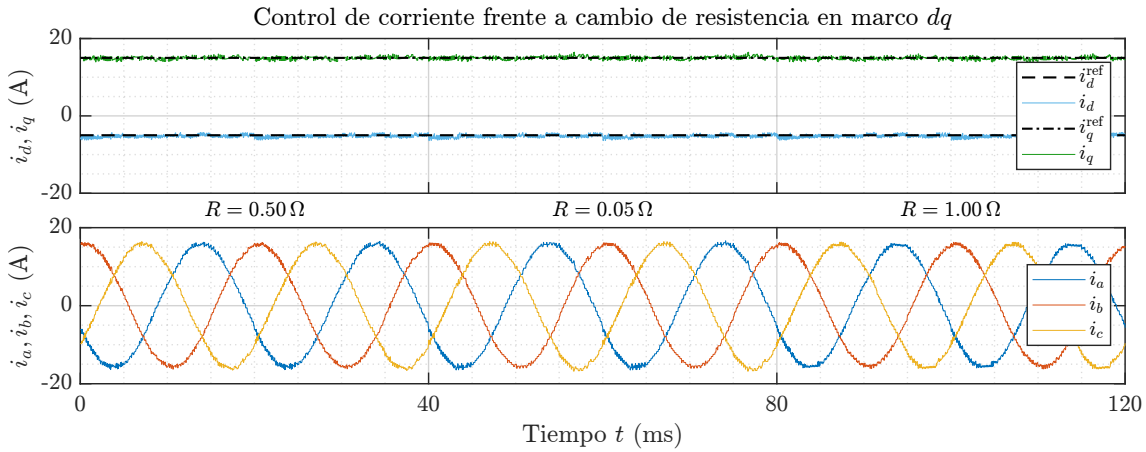


Figura 6.36: Control de corriente en marco de referencia dq frente a cambios en la resistencia de la carga.

A diferencia de lo observado en el caso de cambios de amplitud, no se aprecia acoplamiento transitorio entre las componentes d y q. Esto es consistente con la naturaleza de la perturbación introducida, ya que la variación de la resistencia afecta principalmente la relación tensión–corriente del sistema sin modificar la orientación del vector de corriente en el marco síncrono.

Las corrientes trifásicas mostradas en la parte inferior conservan una forma sinusoidal balanceada en los tres tramos evaluados, sin variaciones visibles en amplitud ni distorsiones adicionales. Este comportamiento confirma que el control basado en DQN en marco dq presenta una baja sensibilidad frente a cambios resistivos, manteniendo un desempeño estable y desacoplado.

6.6.3. Control de corriente frente a variaciones de inductancia de carga

La Figura 6.37 muestra la respuesta del control de corriente en el marco de referencia dq frente a variaciones escalonadas en la inductancia de la carga. El experimento se realiza en régimen estacionario y considera cambios sucesivos del valor inductivo desde $5{,}00$ mH a $2{,}50$ mH, y posteriormente a $7{,}50$ mH, manteniendo constante la amplitud y la frecuencia de la corriente de referencia.

En la parte superior de la figura se observa que las componentes i_d e i_q mantienen un seguimiento adecuado de sus referencias durante todo el experimento. Al reducir la inductancia se aprecia un incremento del rizado en ambas componentes, asociado a la mayor rapidez de la dinámica del sistema y a la disminución del efecto de filtrado natural de la carga. Al aumentar la inductancia, el rizado se atenúa nuevamente, reflejando una respuesta más suavizada.

A diferencia de lo observado en los cambios de amplitud, las variaciones de inductancia no introducen acoplamientos transitorios significativos entre las componentes d y q. Ambas corrientes permanecen próximas a sus valores de referencia, lo que indica que el controlador mantiene el desacople propio del marco dq incluso ante modificaciones en la dinámica eléctrica del sistema.

Las corrientes trifásicas mostradas en la parte inferior conservan un comportamiento balanceado y sinusoidal en los tres escenarios evaluados. Si bien la reducción de la inductancia incrementa el contenido de alta frecuencia en la forma de onda, el control no presenta pérdida de estabilidad ni distorsiones persistentes, confirmando una regulación robusta frente a este tipo de perturbaciones.

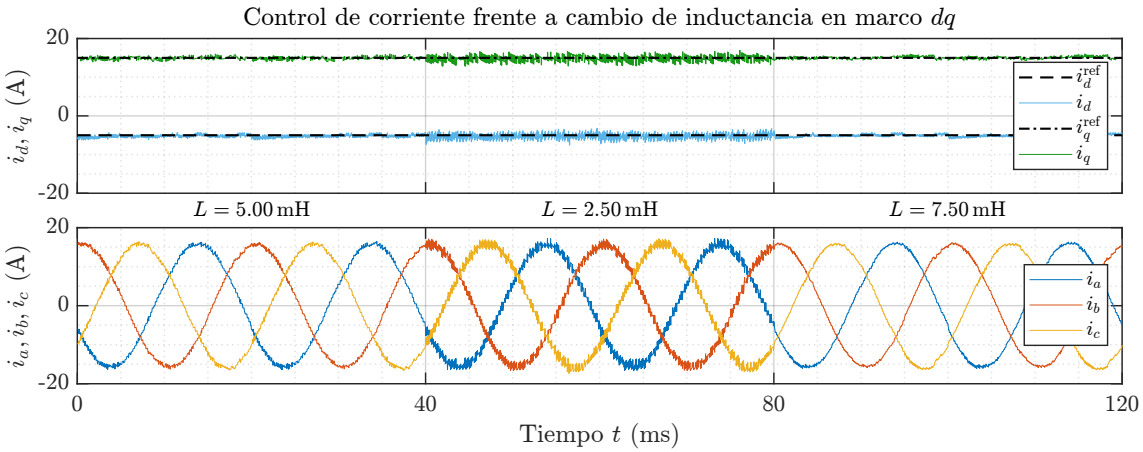


Figura 6.37: Control de corriente en marco de referencia dq frente a cambios en la inductancia de la carga.

6.6.4. Control de corriente frente a variaciones de frecuencia eléctrica

La Figura 6.38 muestra la respuesta del control de corriente en el marco de referencia dq frente a incrementos escalonados en la frecuencia eléctrica de la referencia, desde 50 Hz hasta 300 Hz, manteniendo constante la amplitud de la corriente.

En la parte superior de la figura se observa que, para frecuencias bajas y medias, las componentes i_d e i_q se mantienen cercanas a sus respectivas referencias. Sin embargo, a medida que la frecuencia aumenta, ambas corrientes comienzan a desviarse progresivamente de los valores deseados, aun cuando el sistema continúa generando corrientes trifásicas sinusoidales coherentes con la nueva frecuencia impuesta.

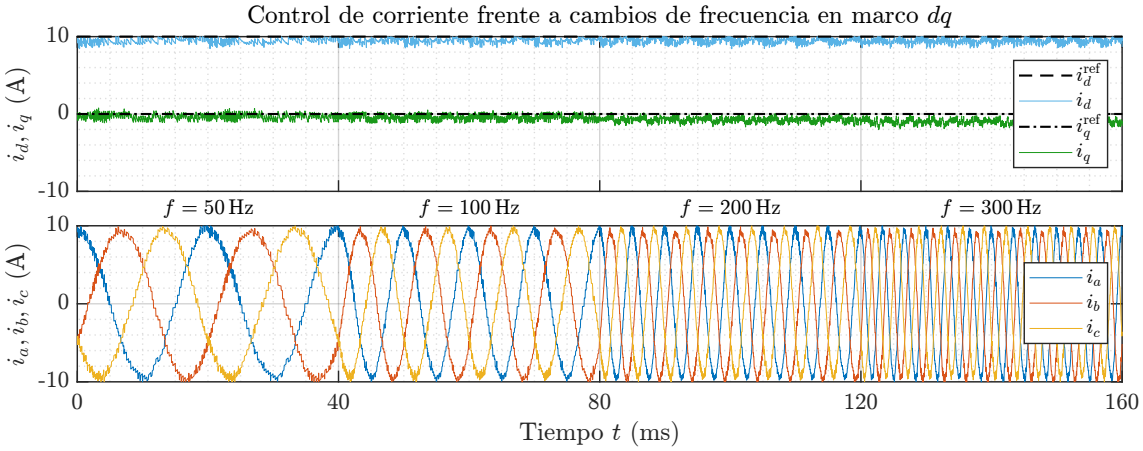


Figura 6.38: Control de corriente en marco de referencia dq frente a cambios en la frecuencia de referencia.

Este comportamiento sugiere que el controlador prioriza mantener la sincronización y la frecuencia de la señal generada, aun a costa de relajar parcialmente el seguimiento exacto de las componentes en el marco dq. Dado que la política de control se implementa mediante una *Look-Up Table*, esta desviación puede atribuirse a las limitaciones inherentes a la discretización del espacio de estados, así como a la ausencia de un mecanismo explícito de adaptación *online* que permita reajustar la política frente a dinámicas más rápidas.

En la parte inferior de la figura se aprecia que, pese a las desviaciones observadas en el marco dq, las corrientes trifásicas permanecen balanceadas y estables durante todo el barrido de frecuencia. Esto indica que el controlador logra sostener la operación en el nuevo régimen dinámico, aunque con una pérdida gradual de precisión en la regulación de las componentes síncronas a medida que se incrementa la frecuencia eléctrica.

Capítulo 7

Control de velocidad en máquina asincrónica

En este capítulo se aborda la extensión del control de corriente basado en aprendizaje por refuerzo al problema de control de velocidad de una máquina asincrónica trifásica. A diferencia de los capítulos anteriores, centrados exclusivamente en el seguimiento de corriente para cargas RL, el objetivo aquí es evaluar el desempeño del controlador en un sistema electromecánico de mayor complejidad, incorporando la dinámica propia del motor y una jerarquía de control adicional.

El esquema de control propuesto se basa en una estructura híbrida que combina elementos del control de campo orientado indirecto (*Indirected Field-Oriented Control*, IFOC) con un controlador de corriente implementado mediante aprendizaje por refuerzo, como se muestra en la Figura 7.1. En particular, se mantiene un lazo externo de control de velocidad convencional, mientras que el lazo interno de control de corriente, típicamente implementado mediante reguladores PI en el marco dq, es reemplazado por un agente Q-learning o Deep Q-Network (DQN).

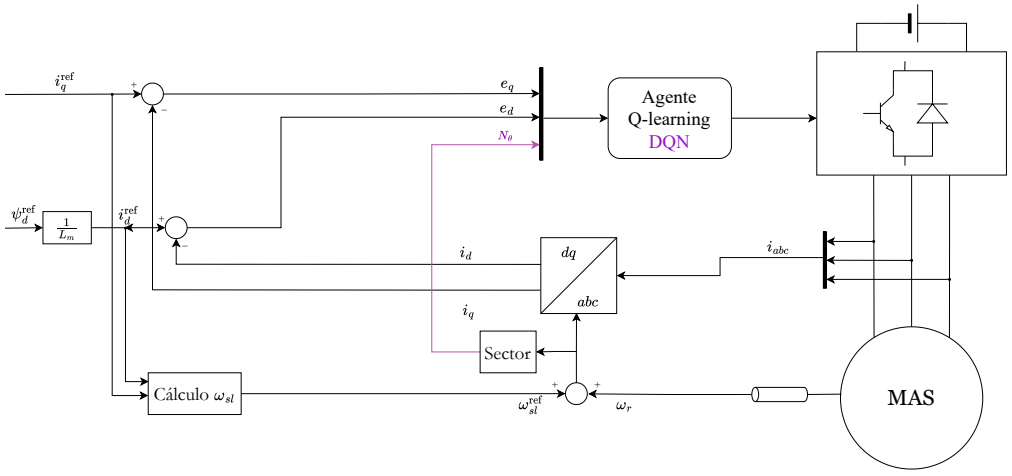


Figura 7.1: Esquema de control propuesto QL-IFOC y DQN-IFOC para motor de inducción trifásico.

Dado que es un control basado en el FOC, a partir de ahora en este texto se nombrará como QL-FOC o QL-IFOC indistintamente para referirse al control híbrido basado en FOC en donde se reemplaza al lazo interno de corrientes por un agente Q-learning. Asimismo, con el caso de

DQN-FOC o DQN-IFOC se referirá al control híbrido basado en FOC en donde se reemplaza al lazo interno de corrientes por un agente Deep Q-Network.

El lazo externo de velocidad utiliza un regulador PI clásico, el cual recibe como entradas la referencia de velocidad mecánica ω_{ref} y la velocidad medida ω_{mec}, generando como salida la referencia de corriente i_q^{ref}. La referencia de corriente directa i_d^{ref} se fija en cero, siguiendo el principio de desacople de flujo del control FOC estándar. De este modo, el control de velocidad conserva la misma estructura de alto nivel que un esquema FOC convencional, permitiendo una comparación directa entre ambos enfoques.

El lazo externo de velocidad utiliza un regulador PI clásico, el cual recibe como entradas la referencia de velocidad mecánica ω_{ref} y la velocidad medida ω_{mec}, generando como salida una referencia de corriente i_q^{ref} asociada al control de par. En paralelo, la referencia i_d^{ref} se define para asegurar la magnetización de la máquina y mantener el nivel de flujo requerido para la conversión electromecánica de energía, pudiendo fijarse como un valor constante nominal o ajustarse según el punto de operación. De este modo, el control de velocidad conserva la estructura jerárquica característica de un esquema FOC convencional, lo que facilita la comparación directa entre el lazo interno clásico basado en PI y su reemplazo por un agente de aprendizaje por refuerzo.

La transformación entre los marcos de referencia dq y abc se realiza mediante las transformaciones inversas correspondientes, utilizando el ángulo eléctrico calculado a partir de la velocidad mecánica. Las acciones generadas por el agente de aprendizaje por refuerzo se traducen directamente en las señales de conmutación del inversor trifásico, cerrando el lazo interno de corriente y permitiendo la regulación del par electromagnético de la máquina.

Cabe destacar que, para el caso del agente DQN, fue necesario realizar un reentrenamiento del controlador de corriente respecto del utilizado en los experimentos con carga RL. En particular, se incrementó la tensión del bus DC desde 200 V a 400 V, con el fin de asegurar coherencia entre el rango de operación del inversor y los valores nominales de la máquina asincrónica considerada. Los parámetros eléctricos y mecánicos de dicha máquina se detallan en el apéndice correspondiente.

A partir de esta estructura de control, el capítulo presenta una serie de experimentos en *PLECS* destinados a evaluar el desempeño del control de velocidad bajo distintas condiciones de operación, incluyendo arranque, seguimiento de referencia e interacción con perturbaciones mecánicas, así como una comparación con un esquema FOC implementado bajo las mismas condiciones.

Finalmente, cabe señalar que los parámetros nominales de la máquina asincrónica considerada, así como los valores base utilizados para la normalización de las variables en unidades por unidad (p.u.), se presentan en el Apéndice E. Esta información permite contextualizar los resultados mostrados en este capítulo y facilita la reproducibilidad de los ensayos realizados.

7.1. Control de velocidad QL-FOC para motor de inducción

7.1.1. Arranque de MAS QL-FOC

La Figura 7.2 muestra el arranque del motor de inducción trifásico bajo el esquema QL-FOC, considerando una referencia de velocidad reducida de $\omega_{\mathrm{ref}} = 0{,}4\,\omega_{\mathrm{nom}}$ y operación sin carga mecánica. Esta condición fue adoptada debido a limitaciones observadas en la capacidad del agente Q-learning para mantener un nivel adecuado de corriente i_q durante aceleraciones hacia velocidades cercanas al valor nominal.

En particular, al intentar realizar un arranque hasta velocidad nominal, el agente no logra sostener de forma consistente la referencia de i_q^{ref} impuesta por el lazo externo de velocidad, lo que deriva en una pérdida progresiva de control del par electromagnético y, finalmente, en la inestabilidad del sistema. Este comportamiento se atribuye a la naturaleza *online* del aprendizaje Q-learning, así como a la discretización limitada del espacio de estados y acciones, que restringe la capacidad del agente para adaptarse a dinámicas más exigentes durante transitorios prolongados.

Por esta razón, todas las pruebas asociadas al esquema QL-FOC, incluyendo el arranque, se realizan a velocidad reducida, lo que permite evaluar el desempeño del control en un régimen donde el agente es capaz de mantener una política estable y reproducible.

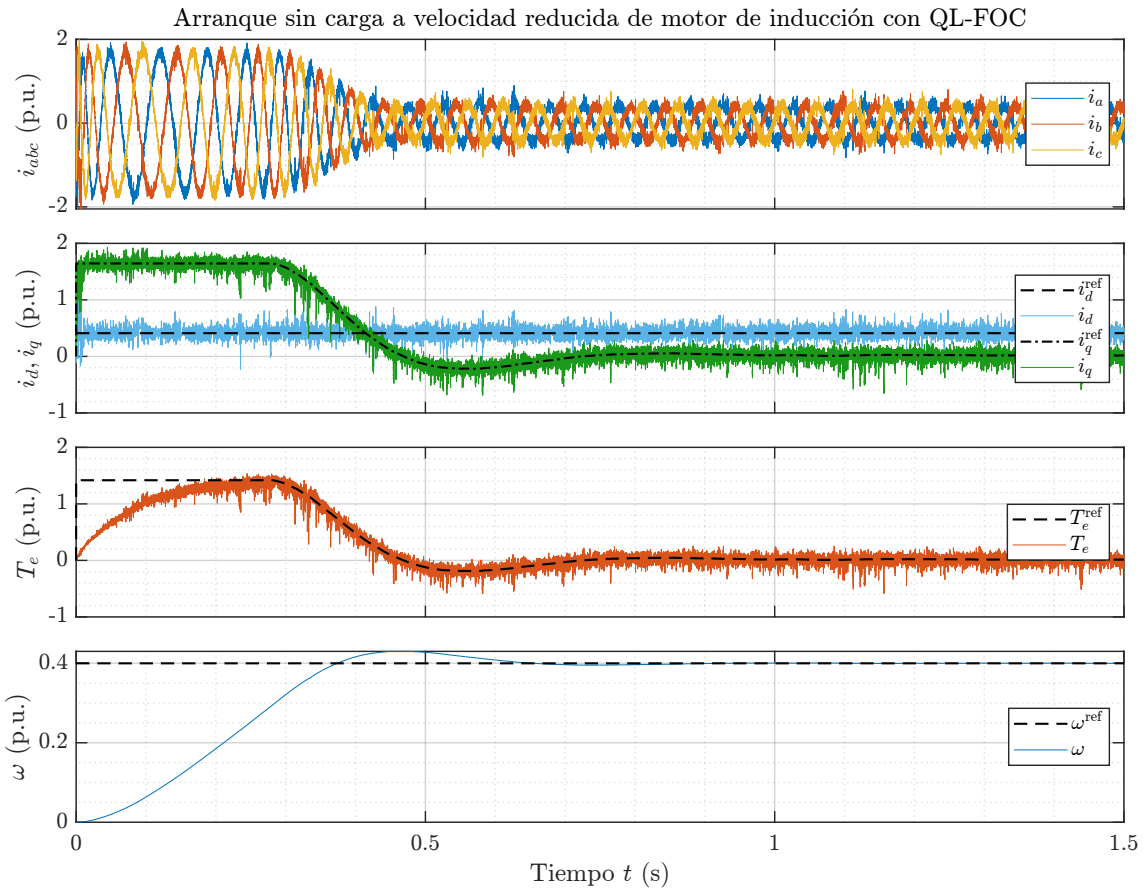


Figura 7.2: Arranque de MAS a velocidad de referencia $\omega_{\text{ref}} = 0.4{\cdot}\omega_{\text{nom}}$ con QL-FOC.

En la figura se observa que la velocidad mecánica converge de manera suave hacia la referencia impuesta, sin oscilaciones sostenidas ni sobrepasos significativos. La corriente i_q presenta un incremento inicial asociado a la generación de par durante la aceleración, mientras que i_d se mantiene en torno al valor de magnetización definido. El par electromagnético refleja este comportamiento, alcanzando un máximo durante el transitorio y estabilizándose una vez alcanzado el régimen permanente.

7.1.2. Seguimiento de referencia de velocidad QL-FOC

La Figura 7.3 muestra la respuesta del esquema QL-FOC ante cambios escalonados en la referencia de velocidad mecánica, considerando operación sin carga mecánica y velocidades dentro del

rango reducido definido para este agente. El experimento se realiza una vez alcanzado el régimen estacionario inicial, aplicando variaciones sucesivas en la referencia ω_{ref}.

En el panel inferior de la figura se observa que la velocidad mecánica sigue la referencia impuesta con un comportamiento estable, presentando transitorios suaves y sin oscilaciones sostenidas. Si bien se aprecia un retardo en la respuesta durante los cambios de consigna, este se mantiene acotado y coherente con la acción del regulador PI del lazo externo de velocidad.

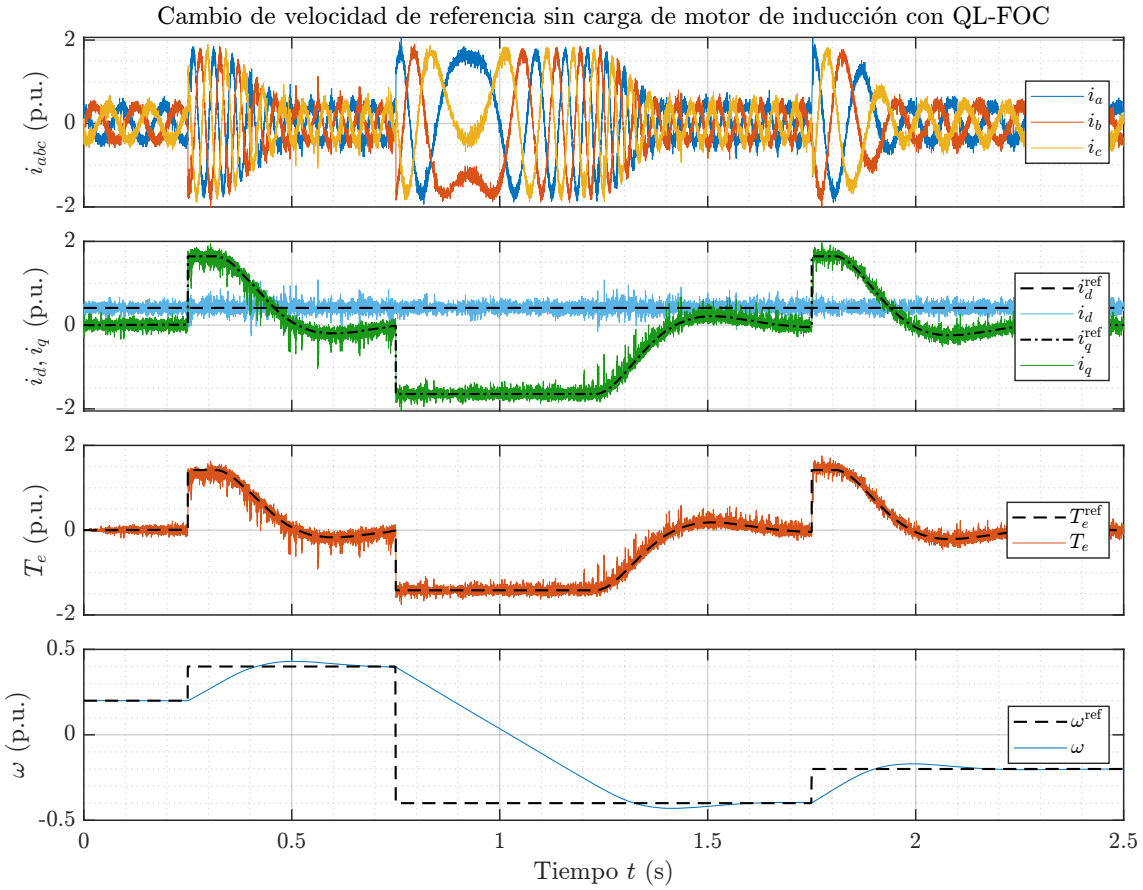


Figura 7.3: Cambio de velocidad de referencia sin carga mecánica con QL-FOC.

El segundo panel muestra la evolución de las corrientes i_d e i_q en p.u. Durante los cambios de velocidad, la componente i_q se incrementa o disminuye de acuerdo con la demanda de par electromagnético impuesta por el regulador de velocidad, mientras que i_d se mantiene en torno al valor de magnetización definido. No obstante, se observa un aumento del rizado y una mayor dispersión en ambas componentes durante los transitorios, reflejando la capacidad limitada del agente Q-learning para sostener una política de control óptima frente a cambios dinámicos prolongados.

En el panel correspondiente al par electromagnético se evidencia una respuesta coherente con la variación de i_q, alcanzando valores máximos durante las aceleraciones y desaceleraciones, y estabilizándose una vez alcanzado el nuevo régimen permanente. Finalmente, las corrientes trifásicas presentadas en el panel superior conservan un comportamiento balanceado durante todo el experimento, sin pérdida de sincronismo entre fases.

Estos resultados muestran que el esquema QL-FOC es capaz de realizar seguimiento de referencia de velocidad bajo condiciones de operación sin carga y a velocidad reducida. Sin embargo, el

incremento del rizado en las variables internas durante los transitorios anticipa las limitaciones del enfoque Q-learning *online* cuando se enfrenta a cambios más exigentes en la dinámica del sistema.

7.1.3. Cambio de carga mecánica QL-FOC

La Figura 7.4 muestra la respuesta del esquema QL-FOC ante variaciones escalonadas del torque de carga mecánica, manteniendo constante la referencia de velocidad en $\omega_{\text{ref}} = 0{,}4\,\omega_{\text{nom}}$. El experimento se inicia en condición sin carga, aplicándose posteriormente incrementos progresivos del torque resistente con el fin de evaluar la capacidad del control para sostener la velocidad frente a perturbaciones mecánicas.

En particular, a $t = 0.5$ s se aplica una carga equivalente al 30 % del torque nominal ($0.3\,T_{\text{nom}}$), a $t = 1.5$ s se incrementa a $0.6\,T_{\text{nom}}$, y finalmente, a partir de $t = 2.5$ s, se aplica el torque nominal completo$T_{\text{nom}} = 11.2\,\text{Nm}$.

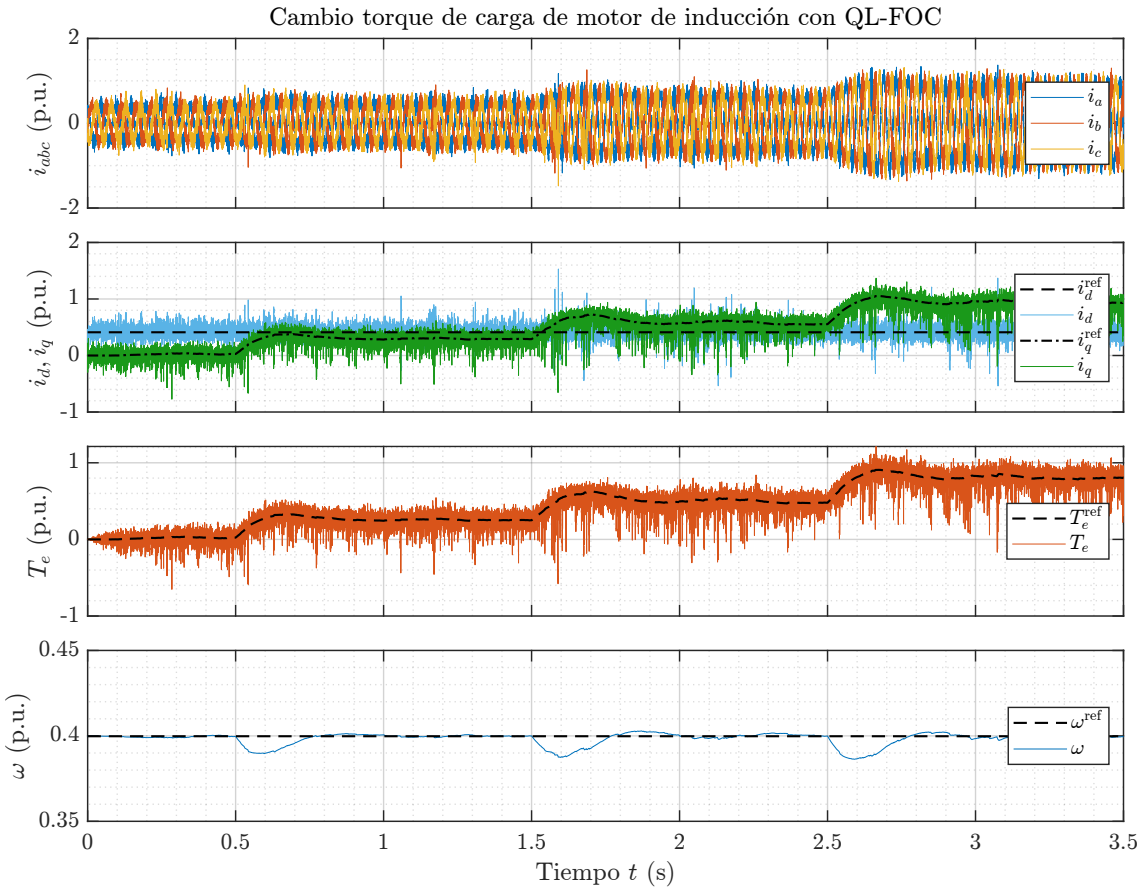


Figura 7.4: Cambio de carga mecánica a velocidad constante con QL-FOC.

En el panel inferior se observa que la velocidad mecánica presenta caídas transitorias en cada aplicación de carga, las cuales son parcialmente compensadas por el regulador PI del lazo externo. Si bien el sistema logra recuperar la velocidad de referencia en todos los casos, el aumento progresivo del torque resistente provoca desviaciones más pronunciadas y una mayor exigencia sobre el control a medida que la carga se aproxima al valor nominal.

El segundo panel muestra la evolución de las corrientes i_d e i_q en p.u. Ante cada incremento de car-

ga, la componente i_q aumenta de forma coherente con la mayor demanda de par electromagnético, mientras que i_d se mantiene en torno al valor de magnetización definido. Sin embargo, se aprecia un aumento del rizado y de la dispersión en ambas componentes, reflejando las limitaciones del agente Q-learning para adaptar su política de control de manera *online* frente a perturbaciones mecánicas sostenidas.

En el panel correspondiente al par electromagnético se observa una respuesta alineada con la evolución de i_q, alcanzando valores crecientes a medida que se incrementa la carga mecánica aplicada. Cabe destacar que el valor de torque nominal T_{nom} utilizado en este experimento no coincide exactamente con el valor base de torque empleado para la normalización en p.u., dado que T_{base} fue calculado a partir de la potencia y velocidad nominales de la máquina, mientras que T_{nom} se determinó a partir de la corriente nominal. Esta diferencia no afecta el análisis cualitativo del desempeño del control, ya que las variaciones de carga se interpretan en términos relativos respecto del torque nominal aplicado.

Las corrientes trifásicas mostradas en el panel superior conservan un comportamiento balanceado durante todo el experimento, sin pérdida de sincronismo entre fases. No obstante, el aumento del rizado asociado a mayores niveles de carga evidencia el incremento en la exigencia dinámica del sistema y anticipa las limitaciones del enfoque Q-learning para aplicaciones de control de velocidad bajo cargas elevadas.

7.2. Control de velocidad DQN-FOC para motor de inducción

En esta sección se presenta el control de velocidad de la máquina asincrónica trifásica utilizando un esquema DQN-FOC, en el cual el lazo interno de control de corriente en el marco dq es implementado mediante un agente *Deep Q-Network*. A diferencia del caso QL-FOC, el agente DQN ejecuta una política fija obtenida durante una etapa previa de entrenamiento, lo que permite evaluar el desempeño del control sin aprendizaje *online* y con un comportamiento reproducible desde el inicio de la simulación.

Para este caso fue necesario realizar un reentrenamiento del controlador de corriente DQN respecto al utilizado en los ensayos con carga RL trifásica. En particular, el valor del bus de continua V_{dc} empleado durante el entrenamiento se incrementó desde $200\ \mathrm{V}$ a $400\ \mathrm{V}$, con el fin de asegurar coherencia con los valores nominales de tensión requeridos por la máquina asincrónica y el inversor utilizado en el esquema de control de velocidad.

Adicionalmente, durante la implementación del controlador en *PLECS*, se observó que el valor de $V_{\mathrm{dc}} = 400\ \mathrm{V}$ no permitía un control adecuado de las corrientes mediante la política almacenada en la *Look-Up Table* (LUT), lo que se manifestaba en una pérdida de estabilidad del lazo interno. Por esta razón, el valor del bus de continua en la simulación final fue ajustado a $V_{\mathrm{dc}} = 460\ \mathrm{V}$, permitiendo que la política DQN opere dentro de un rango de tensión compatible con la dinámica del sistema y mantenga un control estable de las corrientes.

Este ajuste no altera la estructura del esquema DQN-FOC ni la lógica del agente, pero pone de manifiesto la sensibilidad del control basado en LUT al rango operativo para el cual fue entrenada la política, aspecto que resulta relevante en aplicaciones de control jerárquico de máquinas eléctricas.

7.2.1. Arranque de MAS DQN-FOC

La Figura 7.5 muestra el arranque del motor de inducción trifásico bajo el esquema DQN-FOC, considerando operación sin carga mecánica y una referencia de velocidad igual al valor nominal. A diferencia del caso QL-FOC, el controlador basado en DQN permite realizar el arranque hasta velocidad nominal de forma estable, sin pérdida de control del lazo interno de corriente.

En el panel inferior se observa que la velocidad mecánica converge de manera suave hacia la referencia impuesta, sin oscilaciones sostenidas ni sobrepasos relevantes. El par electromagnético presenta un incremento inicial pronunciado durante la aceleración, seguido de una reducción progresiva una vez alcanzado el régimen estacionario, comportamiento coherente con la dinámica esperada del sistema.

El segundo panel muestra la evolución de las corrientes i_d e i_q en p.u. Durante el arranque, la componente i_q se incrementa y permanece saturada en los primeros instantes, proporcionando el par necesario para acelerar la máquina. A medida que la velocidad se aproxima al valor de referencia, el valor medio de i_q tiende a cero, reflejando la disminución de la demanda de par en régimen permanente. Este comportamiento concuerda con la respuesta esperada de un esquema de control de velocidad basado en FOC.

No obstante, se observa que la corriente de referencia i_q^{ref} generada por el regulador PI actúa como una cota superior para la corriente efectiva aplicada por el agente DQN. En particular, durante el transitorio de arranque, i_q^{ref} alcanza valores elevados que no son completamente reproducidos por la corriente medida, aun cuando el agente mantiene una acción de control consistente.

Este fenómeno se atribuye a la forma en que la política del agente DQN es implementada mediante

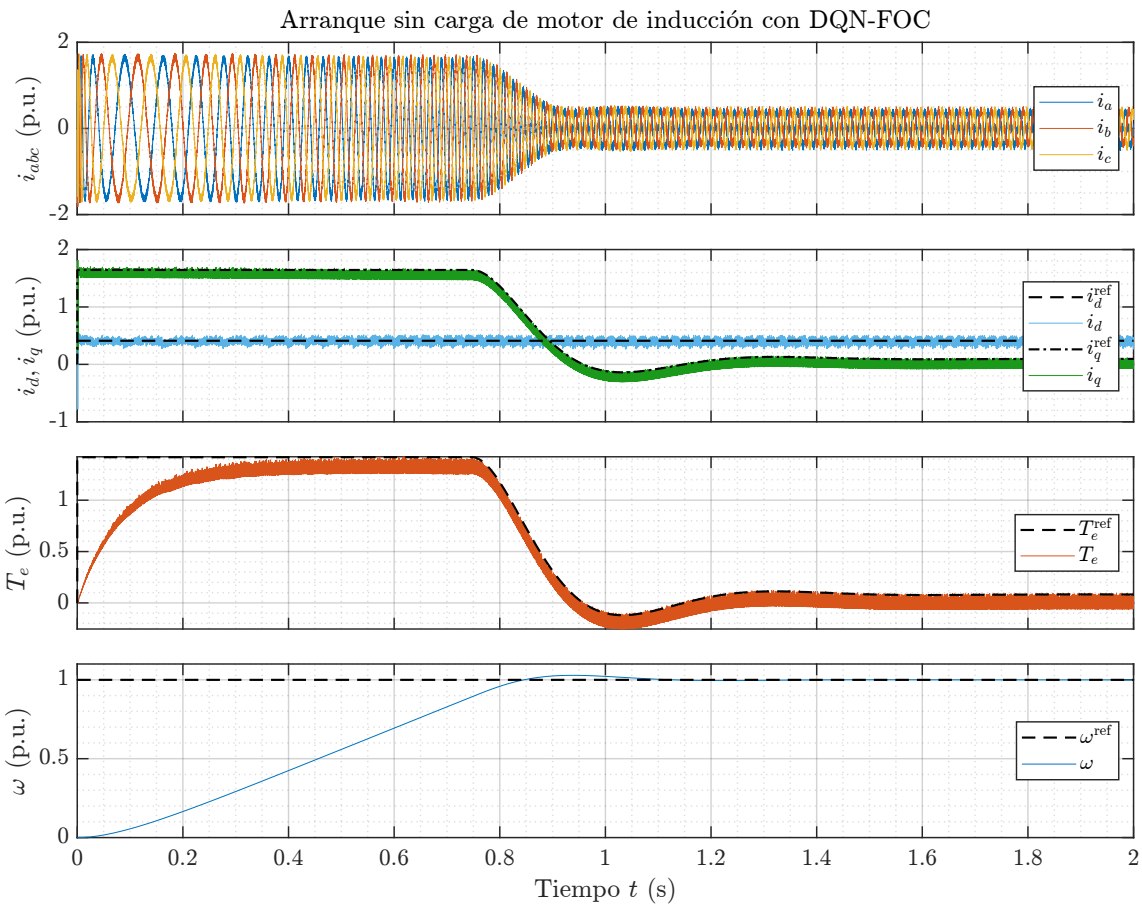


Figura 7.5: Arranque de MAS a velocidad nominal sin carga con DQN-FOC.

una *Look-Up Table*. Dado que la LUT representa una política discretizada y determinista, la acción seleccionada para un determinado par de errores no permite reproducir de manera continua el valor de referencia generado por el regulador PI, sino que establece un límite superior efectivo sobre la corriente que puede inyectarse. Como resultado, la referencia de corriente opera como una señal de saturación implícita, mientras que el agente DQN aplica la acción más cercana disponible dentro de la política aprendida.

Pese a esta limitación, el esquema DQN-FOC mantiene la estabilidad del sistema y logra completar el arranque hasta velocidad nominal, lo que confirma que la política aprendida es suficiente para satisfacer las demandas dinámicas del control de velocidad, aun cuando la referencia de corriente no sea seguida de manera exacta durante los transitorios.

7.2.2. Seguimiento de referencia de velocidad DQN-FOC

La Figura 7.6 muestra la respuesta del esquema DQN-FOC ante cambios escalonados en la referencia de velocidad mecánica, considerando operación sin carga mecánica. El experimento se realiza una vez alcanzado el régimen estacionario inicial, aplicando variaciones sucesivas tanto positivas como negativas en ω_{ref}.

En el panel inferior se observa que la velocidad mecánica sigue adecuadamente las referencias impuestas durante todo el ensayo, incluyendo cambios de signo en la consigna. La respuesta

presenta transitorios suaves, sin oscilaciones sostenidas ni pérdida de estabilidad, evidenciando la capacidad del esquema DQN-FOC para gestionar aceleraciones y desaceleraciones pronunciadas de manera controlada.

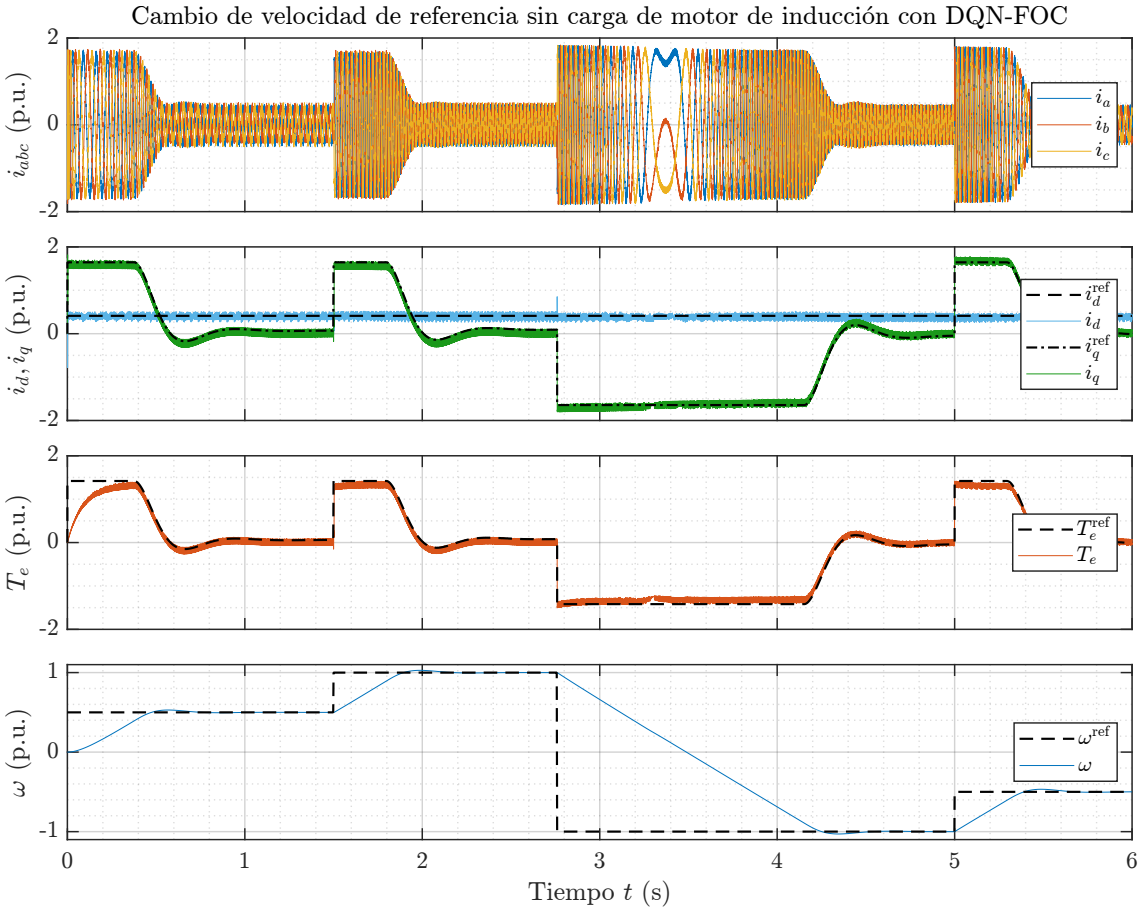


Figura 7.6: Cambio de velocidad de referencia sin carga mecánica con DQN-FOC.

El panel correspondiente al par electromagnético muestra una respuesta coherente con las variaciones de velocidad, incrementándose durante las aceleraciones y cambiando de signo durante las desaceleraciones. La evolución del par se mantiene bien alineada con la referencia generada por el regulador PI del lazo externo, sin saturaciones prolongadas ni comportamientos erráticos.

En el segundo panel se presentan las corrientes i_d e i_q junto con sus referencias. Durante los cambios de velocidad, la componente i_q responde de forma consistente a la demanda de par, mientras que i_d se mantiene en torno al valor de magnetización definido. Al igual que en el arranque, se observa que la referencia de corriente i_q^{ref} actúa como una cota superior, mientras que la corriente efectiva aplicada por el agente DQN se mantiene dentro de los límites impuestos por la política discretizada implementada en la LUT.

Las corrientes trifásicas mostradas en el panel superior conservan un comportamiento balanceado durante todo el experimento, sin pérdida de sincronismo entre fases ni distorsiones abruptas asociadas a los cambios de referencia. Incluso durante transitorios exigentes, la forma de onda se mantiene estable, reflejando la correcta coordinación entre el lazo externo de velocidad y el lazo interno de corriente basado en DQN.

Estos resultados confirman que el esquema DQN-FOC es capaz de realizar seguimiento de refe-

rencia de velocidad de forma robusta y estable en un amplio rango de operación, superando las limitaciones observadas en el enfoque QL-FOC y permitiendo cambios dinámicos de consigna sin degradar el desempeño del control.

7.2.3. Cambio de carga mecánica DQN-FOC

La Figura 7.7 muestra la respuesta del esquema DQN-FOC ante variaciones escalonadas del par de carga mecánica, manteniendo constante la referencia de velocidad. El ensayo se realiza partiendo desde operación sin carga y aplicando incrementos sucesivos del par resistente, permitiendo evaluar la capacidad del controlador para rechazar perturbaciones mecánicas sin comprometer la regulación de velocidad.

En el panel inferior se observa que la velocidad mecánica presenta caídas transitorias acotadas tras cada incremento de carga, recuperando rápidamente el valor de referencia. Estas desviaciones son pequeñas y bien amortiguadas, lo que indica un adecuado desacople entre el lazo de velocidad y las perturbaciones externas introducidas por el par de carga.

El panel correspondiente al par electromagnético muestra un aumento progresivo de T_e en respuesta a los cambios de carga, siguiendo de forma consistente la referencia generada por el regulador PI. La respuesta del par es suave y estable, sin oscilaciones sostenidas ni saturaciones prolongadas, incluso para los niveles de carga más elevados considerados.

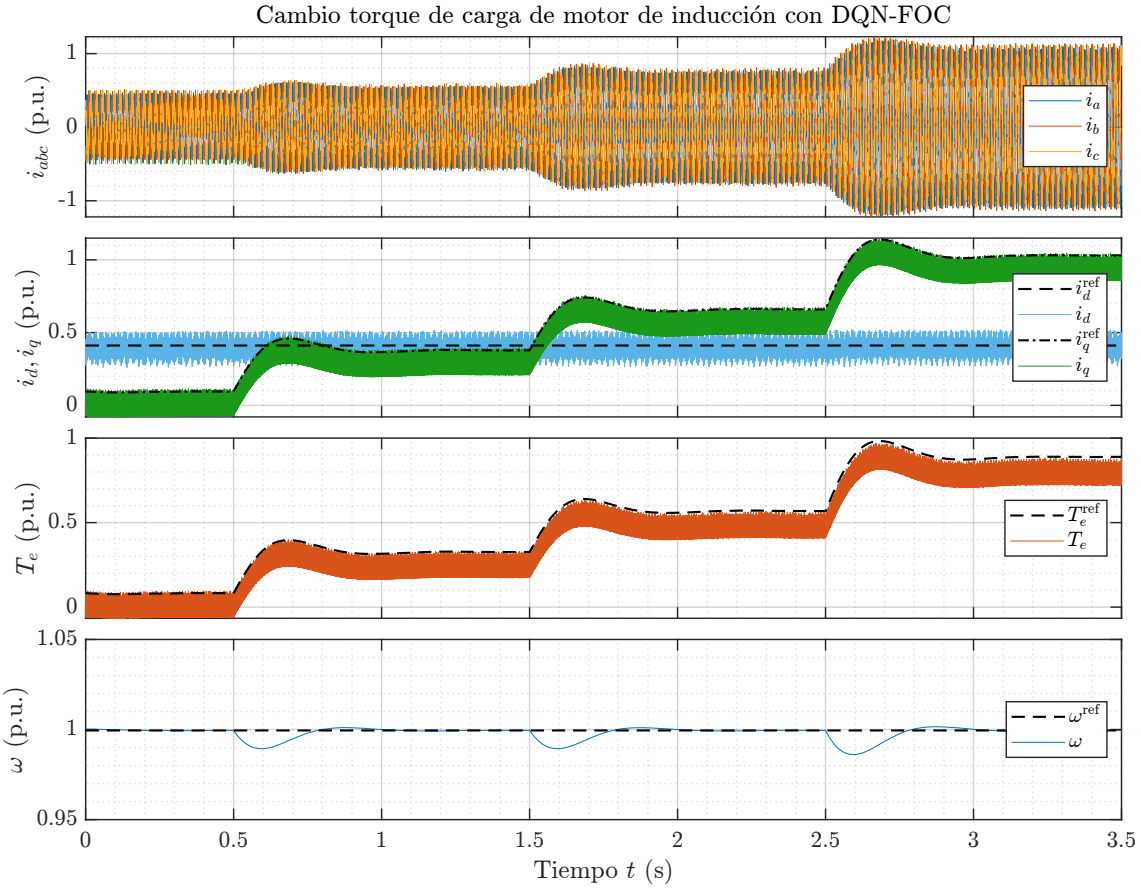


Figura 7.7: Cambio de carga mecánica a velocidad constante con DQN-FOC.

En el segundo panel se presentan las corrientes i_d e i_q junto con sus respectivas referencias. Ante

cada incremento de carga, la componente i_q aumenta para compensar la mayor demanda de par, mientras que i_d se mantiene cercana al valor de magnetización definido. Al igual que en los casos anteriores, se aprecia que la referencia de corriente i_q^{ref} actúa como una cota superior, mientras que la corriente efectiva aplicada por el agente DQN permanece dentro de los límites impuestos por la política discretizada almacenada en la LUT.

Las corrientes trifásicas mostradas en el panel superior conservan un comportamiento balanceado durante todo el ensayo, sin pérdida de simetría ni distorsiones abruptas asociadas a los cambios de carga. A medida que el par resistente aumenta, se observa un incremento en la amplitud de las corrientes, coherente con el mayor esfuerzo electromecánico requerido por el motor.

Estos resultados confirman que el esquema DQN-FOC es capaz de rechazar perturbaciones mecánicas de manera efectiva, manteniendo la velocidad regulada y una respuesta de corriente estable. El controlador demuestra un comportamiento robusto frente a incrementos significativos de carga, sin requerir ajustes adicionales en la política de control implementada.

7.3. Comparación de controles con IFOC

7.3.1. Tensión aplicada por inversor

La Figura 7.8 presenta una comparación entre la tensión de línea v_{ab} aplicada al motor para tres esquemas de control: DQN-FOC con política reentrenada, DQN-FOC utilizando una LUT entrenada a $V_{dc} = 200$ V sin reentrenamiento, y un controlador FOC. Esta comparación permite analizar directamente el esfuerzo de control requerido por cada estrategia y la coherencia entre la política implementada y el sistema físico.

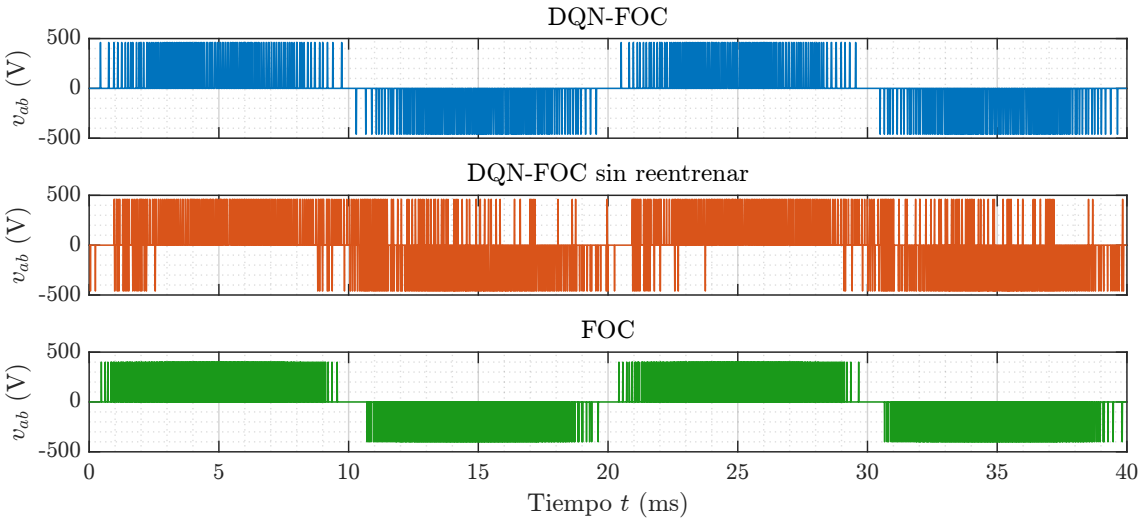


Figura 7.8: Comparación en tensión aplicada en estado estacionario sin carga entre FOC y DQN-FOC.

En el caso del DQN-FOC reentrenado, la tensión aplicada utiliza de manera consistente el rango disponible del fuente DC, presentando un patrón de conmutación estructurado y coherente con la dinámica del motor. Esto indica que la política almacenada en la LUT es compatible con los niveles de tensión reales del inversor y capaz de generar las acciones necesarias para el control de corriente y torque.

Por el contrario, al emplear una LUT entrenada a un valor de V_{dc} inferior, el controlador DQN-FOC

sin reentrenamiento exhibe un comportamiento errático en la tensión aplicada, con conmutaciones desordenadas y saturaciones frecuentes. Este efecto no responde a una inestabilidad del algoritmo, sino a una incongruencia entre el entorno de entrenamiento y el entorno de implementación, que el agente no puede compensar mediante su política fija.

Finalmente, el controlador FOC presenta una señal de tensión más regular y predecible, propia de un esquema basado en modelos analíticos. Este comportamiento sirve como referencia para contrastar la naturaleza del esfuerzo de control generado por los enfoques basados en aprendizaje por refuerzo.

Es importante volver a mencionar que el control DQN-IFOC selecciona directamente los vectores a partir de su LUT, decisión tomada a partir de la lectura de los errores de corrientes en el plano dq y el sector angular en el que se encuentra, en contraste con el IFOC utiliza un SV-PWM para generar los pulsos de disparos a partir de una señal triangular con una frecuencia predefinida.

La Figura 7.9 muestra la comparación de la tensión aplicada al motor en régimen estacionario sin carga mecánica, operando a velocidad reducida, para los esquemas QL-FOC y FOC. En ambos casos se representa la tensión de fase generada por el inversor, lo que permite analizar directamente el esfuerzo de control requerido por cada estrategia.

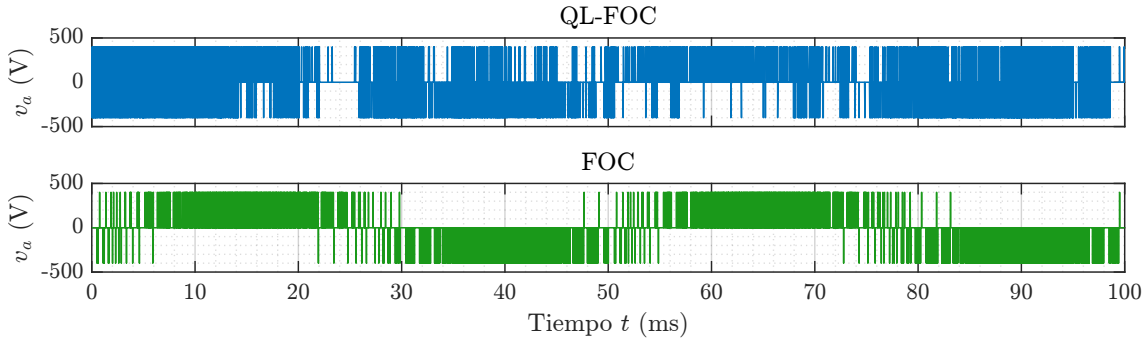


Figura 7.9: Comparación en tensión aplicada en estado estacionario sin carga entre FOC y QL-FOC.

En el caso del control FOC se observa una señal de tensión más estructurada y regular, con patrones de conmutación consistentes y una ocupación uniforme del bus de continua. Este comportamiento es coherente con el uso de reguladores PI lineales y moduladores PWM convencionales, los cuales generan una acción de control suavizada y predecible en régimen estacionario.

Por el contrario, el esquema QL-FOC presenta una señal de tensión con mayor dispersión temporal y una conmutación más irregular. Esta característica se asocia a la naturaleza discreta de la política aprendida por el agente Q-learning, que selecciona acciones de conmutación directamente a partir del estado del sistema sin una etapa explícita de modulación. Como resultado, el control logra mantener el seguimiento de corriente y velocidad a costa de un mayor nivel de variabilidad en la tensión aplicada.

A pesar de estas diferencias, ambos esquemas permiten sostener el régimen estacionario a velocidad reducida sin pérdida de estabilidad. Sin embargo, el mayor esfuerzo de control observado en QL-FOC anticipa un incremento potencial en pérdidas por conmutación y estrés del inversor en comparación con el FOC, aspecto que resulta relevante al evaluar la viabilidad práctica de estrategias basadas en aprendizaje por refuerzo.

7.3.2. Corrientes inyectadas por inversor

La Figura 7.10 muestra la comparación de la corriente de fase i_a entregada por el inversor para los mismos tres esquemas de control analizados previamente: DQN-FOC con política reentrenada, DQN-FOC utilizando una LUT entrenada a $V_{dc} = 200$ V sin reentrenamiento, y el FOC. Dado que se utiliza la misma ventana temporal que en la Figura 7.8, esta comparación permite observar cómo las diferencias en la tensión aplicada se reflejan directamente en la forma de onda de la corriente.

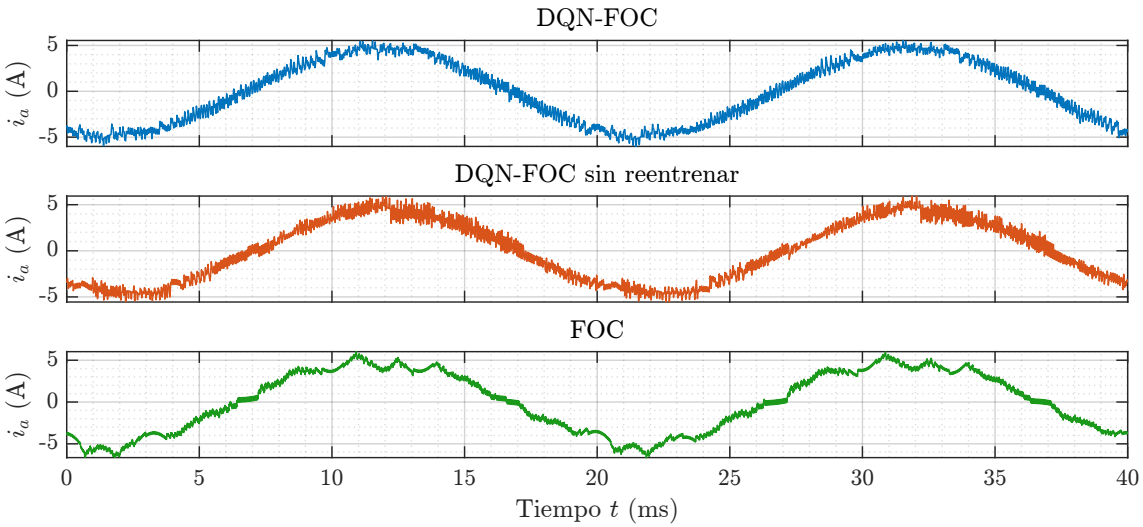


Figura 7.10: Comparación en corriente inyectada por inversor en estado estacionario sin carga entre FOC y DQN-FOC.

En el caso del controlador DQN-IFOC reentrenado, la corriente presenta una forma aproximadamente sinusoidal, con una amplitud y periodicidad coherentes con el régimen estacionario del motor sin carga. Se observa un rizado de alta frecuencia superpuesto a la componente fundamental, el cual es consistente con la naturaleza discreta de las acciones del agente, que selecciona directamente los vectores de tensión del inversor. A pesar de este rizado, la forma global de la corriente se mantiene estable, lo que indica que la política aprendida es capaz de regular adecuadamente la corriente del estator.

Para el caso del DQN-IFOC sin reentrenamiento, la corriente mantiene en términos generales una envolvente sinusoidal similar, pero presenta un nivel de rizado considerablemente mayor. Este comportamiento es consecuencia directa de las conmutaciones erráticas observadas previamente en la tensión aplicada. La política almacenada en la LUT fue entrenada bajo un nivel de tensión distinto al presente en el sistema, lo que provoca decisiones de control subóptimas que se traducen en una mayor variabilidad instantánea de la corriente.

Finalmente, el controlador FOC presenta oscilaciones más visibles alrededor de la componente fundamental de la corriente. Este comportamiento se relaciona con el esquema de modulación utilizado para generar las señales de disparo del inversor, el cual difiere del mecanismo de selección directa de vectores empleado en el DQN-IFOC. En el caso del controlador basado en aprendizaje por refuerzo, el proceso de entrenamiento tiene como objetivo principal el seguimiento de la corriente sinusoidal de referencia, lo que orienta la política aprendida hacia la minimización del error instantáneo de corriente. En contraste, en el esquema IFOC la tensión de referencia se implementa mediante modulación SV-PWM a una frecuencia de conmutación predefinida, lo que

introduce oscilaciones de alta frecuencia en la corriente alrededor de su componente fundamental.

7.3.3. Torque electromagnético producido sobre el motor

En la condición de operación sin carga, el torque electromagnético promedio del motor se mantiene cercano a cero en los tres esquemas de control, tal como muestra la Figura 7.11, lo cual es consistente con el hecho de que el motor únicamente debe compensar las pérdidas internas del sistema. En consecuencia, las diferencias entre los métodos se manifiestan principalmente en el nivel de rizado del torque, para el caso de operación sin carga.

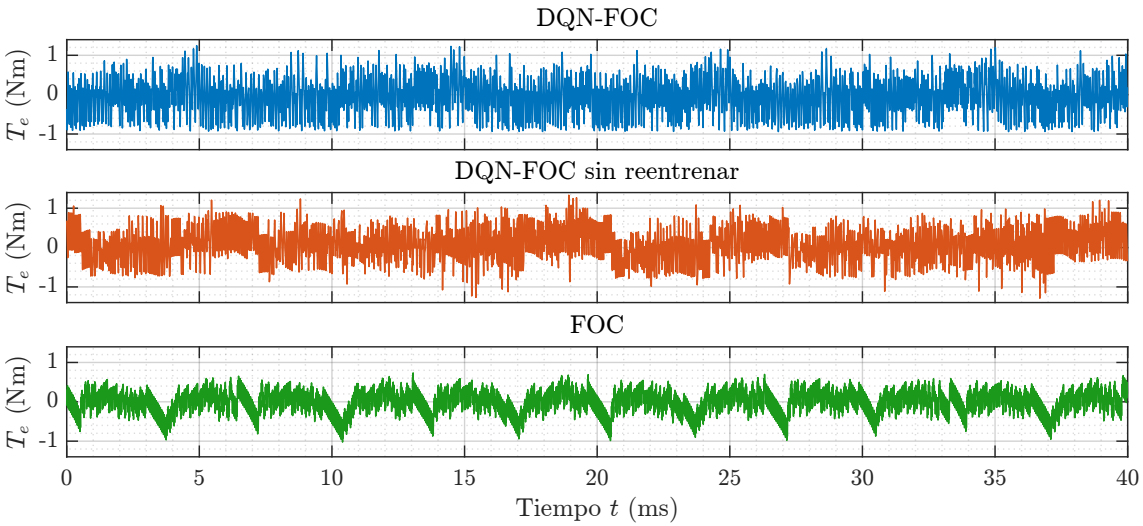


Figura 7.11: Comparación en torque electromagnético generado por el motor en estado estacionario sin carga entre FOC y DQN-FOC.

En el caso del controlador DQN-FOC reentrenado, el torque oscila alrededor de cero con una variabilidad relativamente acotada. Este comportamiento es coherente con la forma de onda de corriente observada previamente, donde el seguimiento de la componente sinusoidal se mantiene estable pese al rizado introducido por la naturaleza discreta de las acciones del agente.

Para el DQN-FOC sin reentrenamiento, el torque presenta una mayor dispersión instantánea y episodios de variación más pronunciados. Este comportamiento está directamente relacionado con las conmutaciones menos estructuradas observadas en la tensión aplicada y con el mayor rizado presente en la corriente del estator, consecuencia de la discrepancia entre las condiciones de entrenamiento del agente y las condiciones reales de operación.

Finalmente, el controlador FOC presenta un patrón de rizado más estructurado y periódico. Este comportamiento se asocia al proceso de modulación empleado en el inversor, donde la tensión de referencia generada por los reguladores se implementa mediante SV-PWM a una frecuencia de conmutación fija. Como resultado, el torque presenta oscilaciones que siguen un patrón repetitivo vinculado al proceso de modulación, la dinámica del inversor y motor.

7.3.4. Comparación de métricas de distintos métodos

En la Tabla 7.1 se presentan las métricas de desempeño obtenidas en condición de operación sin carga mecánica para los esquemas QL–FOC, DQN–FOC y FOC, considerando dos configuraciones de frecuencia de PWM para este último.

Tabla 7.1: Comparación de métricas en operación sin carga.

Control	**THD**$_{i_a}$ **[%]**	**RMSE**$_{i_d}$ **[A]**	**RMSE**$_{i_q}$ **[A]**	f_{sw} **[kHz]**
QL–FOC	17.53	0.983	1.161	20.05
DQN–FOC	4.84	0.458	1.237	8.68
FOC ($f_{\text{pwm}} = 40$ kHz)	4.30	0.672	0.300	28.98
FOC ($f_{\text{pwm}} = 10$ kHz)	3.14	0.372	0.515	10.00

Se observa que el menor nivel de distorsión armónica total corresponde al control FOC operando con $f_{\text{pwm}} = 10$ kHz, alcanzando un THD de $3{,}14\,\%$. Este valor es inferior tanto al obtenido con FOC a 40 kHz ($4{,}30\,\%$) como al del controlador DQN–FOC ($4{,}84\,\%$), lo que indica que, en esta condición de operación, una mayor frecuencia de PWM no garantiza una reducción adicional de la distorsión armónica.

En términos de error de seguimiento, el FOC a 40 kHz presenta el menor RMSE en la componente de par i_q ($0{,}300$ A), mientras que al reducir la frecuencia a 10 kHz dicho error aumenta a $0{,}515$ A. Por el contrario, el RMSE en la componente directa i_d disminuye al reducir la frecuencia de conmutación, lo que evidencia un compromiso entre precisión en el control de flujo y control de par.

El controlador DQN–FOC exhibe valores intermedios de THD y RMSE. En particular, presenta un THD mayor que el FOC a 10 kHz, pero comparable al FOC a 40 kHz, operando con una frecuencia de conmutación significativamente menor ($8{,}68$ kHz). Sin embargo, el error en la componente i_q es superior al del FOC, lo que refleja una menor precisión en el control de par en régimen sin carga.

Finalmente, el esquema QL–FOC presenta los mayores valores de distorsión armónica y error de seguimiento en ambas componentes de corriente, como se aprecia en la Tabla 7.1, evidenciando limitaciones relevantes en la calidad de la forma de onda y en la precisión del control.

La Tabla 7.2 resume las métricas de desempeño obtenidas bajo condición de carga nominal para los distintos esquemas de control evaluados.

Tabla 7.2: Comparación de métricas en operación con carga nominal.

Control	**THD**$_{i_a}$ **[%]**	**RMSE**$_{i_d}$ **[A]**	**RMSE**$_{i_q}$ **[A]**	f_{sw} **[kHz]**
QL–FOC	8.49	1.205	1.646	19.53
DQN–FOC	1.68	0.436	1.277	9.24
FOC ($f_{\text{pwm}} = 40$ kHz)	1.76	0.581	0.417	26.83
FOC ($f_{\text{pwm}} = 10$ kHz)	1.14	0.358	0.529	10.00

En este caso, el menor valor de distorsión armónica total corresponde nuevamente al FOC operando con $f_{\text{pwm}} = 10$ kHz, alcanzando un THD de $1{,}14\,\%$. Este resultado es inferior tanto al FOC a 40 kHz ($1{,}76\,\%$) como al controlador DQN–FOC ($1{,}68\,\%$), confirmando que, bajo carga, la reducción de la frecuencia de conmutación no deteriora la calidad espectral de la corriente y puede incluso mejorarla.

En términos de error de seguimiento, el FOC mantiene los menores valores de RMSE en la componente de par i_q, particularmente a 40 kHz. Al reducir la frecuencia de PWM a 10 kHz, el error en i_q aumenta, mientras que el error en la componente directa i_d disminuye, reproduciendo el mismo compromiso observado en la condición sin carga.

El controlador DQN–FOC presenta un nivel de distorsión armónica bajo y cercano al del FOC, con una frecuencia de conmutación considerablemente menor ($9{,}24$ kHz). No obstante, el RMSE en la componente i_q es superior al del FOC convencional, lo que indica una menor precisión en

el control de par bajo carga, aun cuando la calidad de la forma de onda se mantiene en niveles competitivos.

El esquema QL–FOC vuelve a exhibir los peores indicadores de desempeño en todas las métricas consideradas, con valores elevados de THD y error de seguimiento, lo que limita su aplicabilidad práctica en esquemas de control de velocidad de máquinas eléctricas bajo carga.

7.4. Comentarios finales de control de velocidad

En este capítulo se evaluó la extensión de los controladores de corriente basados en aprendizaje por refuerzo a un esquema de control de velocidad de una máquina asincrónica, integrándolos como reemplazo del lazo interno de corrientes dq en un control FOC convencional. Los resultados obtenidos permiten constatar que tanto el enfoque QL–FOC como el DQN–FOC son capaces de operar dentro de una arquitectura de control de velocidad clásica, manteniendo estabilidad y seguimiento adecuado en distintas condiciones de operación.

El análisis comparativo evidencia que el controlador DQN–FOC logra un compromiso favorable entre calidad de corriente y esfuerzo de conmutación, alcanzando niveles de distorsión armónica cercanos a los del FOC, pero con una frecuencia de conmutación significativamente menor. No obstante, el FOC convencional continúa presentando la mejor precisión en el control de par, especialmente en términos del error de seguimiento de la componente i_q, lo que se mantiene como una ventaja relevante en aplicaciones donde la respuesta dinámica del par es prioritaria.

Por su parte, el esquema QL–FOC muestra mayores limitaciones al ser integrado en el control de velocidad, particularmente en condiciones de mayor exigencia dinámica, lo que se refleja en mayores errores de seguimiento y una menor robustez frente a variaciones de carga y referencia. En conjunto, estos resultados confirman que los controladores basados en aprendizaje por refuerzo pueden integrarse de manera efectiva en estructuras de control de velocidad, destacando al enfoque DQN como una alternativa viable al control FOC cuando se prioriza la reducción del esfuerzo de conmutación, sin comprometer de forma significativa la calidad del control.

Capítulo 8

Conclusiones

En este trabajo se abordó el diseño, entrenamiento e implementación de estrategias de control de corriente y velocidad basadas en aprendizaje por refuerzo, utilizando algoritmos Q-learning y Deep Q-Network (DQN), aplicados inicialmente a cargas RL y posteriormente integrados en un esquema de control de velocidad de una máquina de inducción trifásica. A lo largo de la tesis se analizaron tanto configuraciones monofásicas como trifásicas, considerando distintos marcos de referencia, discretizaciones y plataformas de simulación.

Control de corriente monofásico

En el caso del control de corriente monofásico mediante Q-learning, se demostró que el agente es capaz de aprender una política de control efectiva de manera *online*, logrando un seguimiento adecuado de referencias sinusoidales dentro de los límites físicos del sistema. El análisis univariado de hiperparámetros evidenció una marcada sensibilidad del desempeño a la elección de la tasa de aprendizaje, el factor de descuento y la discretización del error de corriente. En particular, la utilización de la mediana como métrica representativa permitió caracterizar de forma más robusta el comportamiento típico del agente frente a la variabilidad inherente al proceso de aprendizaje. No obstante, el enfoque tabular presentó limitaciones asociadas al compromiso entre resolución del espacio de estados y tamaño de la tabla Q, lo que restringe su escalabilidad.

El control monofásico mediante DQN mostró una mejora sustancial respecto al Q-learning tabular, tanto en términos de distorsión armónica total como de suavidad en la acción de control. El entrenamiento *offline* permitió desacoplar el proceso de aprendizaje de la ejecución en línea, posibilitando la exportación de la política aprendida mediante una *Look-Up Table*. Los resultados en PLECS evidenciaron un seguimiento de corriente más estable y continuo, incluso frente a variaciones abruptas de los parámetros de la carga. Asimismo, se observó que la resolución de la LUT y su rango de construcción influyen directamente en la capacidad del controlador para manejar cambios de amplitud y frecuencia, destacando la importancia del diseño de la discretización en este tipo de implementaciones.

Control de corriente trifásico

En los casos trifásicos, tanto Q-learning como DQN fueron evaluados en los marcos de referencia $\alpha\beta$ y dq. Para el Q-learning trifásico en $\alpha\beta$, el agente demostró la capacidad de aprender una

política de control directamente desde una tabla Q inicializada en cero, logrando estabilizar el sistema tras una fase transitoria de aprendizaje. Sin embargo, al extender el enfoque al marco dq, se evidenció la necesidad de utilizar tiempos de muestreo significativamente menores y reducir la dimensionalidad de la tabla Q, reflejando el aumento en la exigencia dinámica y la sensibilidad del sistema a la discretización temporal.

El control trifásico mediante DQN permitió superar estas limitaciones, particularmente en el marco $\alpha\beta$, donde el controlador mantuvo un desempeño estable con tiempos de muestreo del orden de decenas de microsegundos. En el caso del DQN en dq, la inclusión del sector angular como parte de las observaciones resultó clave para lograr una correcta generalización de la política, evidenciando la importancia de incorporar información contextual que desambigüe la acción óptima en sistemas periódicos. En conjunto, los resultados muestran que el uso de DQN, junto con una adecuada selección de observaciones y discretización, ofrece una alternativa viable y escalable para el control de corriente trifásico basado en aprendizaje por refuerzo.

Control de velocidad de máquina de inducción

Finalmente, se integraron los controladores de corriente basados en Q-learning y DQN dentro de un esquema de control de velocidad tipo FOC para una máquina de inducción trifásica. En este contexto, el lazo externo de velocidad se mantuvo como un regulador PI clásico, mientras que el lazo interno de corrientes fue reemplazado por los agentes de aprendizaje. Para el caso QL–FOC, se observó que el controlador presenta limitaciones para mantener la corriente de par durante arranques a velocidad nominal, lo que obligó a realizar las pruebas a velocidades reducidas. Aun así, el esquema permitió analizar el comportamiento del agente frente a variaciones de carga mecánica y referencias de velocidad.

El enfoque DQN–FOC mostró un desempeño más robusto, logrando arranques estables y un seguimiento adecuado de la velocidad bajo distintas condiciones de carga, siempre que la política fuera entrenada con parámetros coherentes con los valores nominales del sistema. Los resultados comparativos con un FOC clásico evidenciaron que el control basado en DQN puede alcanzar niveles de distorsión armónica y error de seguimiento comparables, o incluso superiores en ciertos regímenes, utilizando frecuencias de conmutación significativamente menores. Este compromiso favorable entre calidad de control y esfuerzo de conmutación destaca el potencial del aprendizaje por refuerzo profundo como herramienta para el diseño de estrategias de control avanzadas en sistemas eléctricos.

Conclusión general

En conjunto, los resultados obtenidos confirman que los métodos de aprendizaje por refuerzo, particularmente aquellos basados en DQN, constituyen una alternativa viable y flexible para el control de corriente y velocidad en sistemas eléctricos de potencia. Si bien el Q-learning tabular permite estudiar en detalle el proceso de aprendizaje y sus dependencias paramétricas, su aplicabilidad práctica se ve limitada por cuestiones de escalabilidad y requerimientos temporales. En contraste, el uso de DQN junto con la exportación de políticas mediante LUT permite combinar las ventajas del aprendizaje profundo con la ejecución determinista requerida en aplicaciones de control en tiempo real, abriendo la puerta a futuras extensiones hacia sistemas de mayor complejidad y aplicaciones industriales.

Apéndice A

Mecanismos adicionales de estabilización de la política de acción

El comportamiento del agente de control puede verse influido por la forma en que decide y mantiene las acciones de conmutación. Durante las primeras simulaciones se observó que, aunque el seguimiento de la referencia era adecuado, en ciertos escenarios aparecían oscilaciones lentas, incrementos en la frecuencia de conmutación o fluctuaciones no deseadas en la corriente. Estos efectos se intensificaban cuando el agente estaba autorizado a cambiar de acción en cada paso de simulación, lo que motivó la incorporación de mecanismos complementarios orientados a suavizar la política y mejorar la estabilidad del control.

A continuación se describen los mecanismos utilizados:

- **Action Hold (AH):** establece el número mínimo de pasos de simulación que deben transcurrir antes de permitir una nueva decisión del agente. Al restringir la frecuencia de actualización de la acción, este mecanismo reduce cambios excesivamente rápidos, favoreciendo una dinámica más estable, aunque a costa de un posible aumento del rizado [26].
- **Action Sticky (AS):** introduce una probabilidad de repetir la acción ejecutada en el paso anterior. Esta correlación temporal entre acciones consecutivas actúa como un filtro sobre la política aprendida, atenuando variaciones abruptas y disminuyendo la sensibilidad a perturbaciones [26].
- Q_{thresh} **(QT):** define un umbral para la diferencia mínima entre valores Q necesaria para considerar que una acción es significativamente mejor que otra. Cuando las ventajas relativas son pequeñas, el agente mantiene la acción previa, evitando alternancias no deseadas. Este mecanismo se relaciona con estrategias de *thresholded action selection* propuestas en la literatura [27].

Estos mecanismos no forman parte del algoritmo Q-learning estándar, pero su inclusión permitió mejorar la estabilidad del sistema en condiciones donde la dinámica eléctrica es rápida y la toma de decisiones discreta puede amplificar efectos no deseados. Su uso se justifica en los capítulos de resultados, donde se muestran los escenarios específicos en los que contribuyen a obtener un comportamiento más robusto.

Apéndice B

Resultados complementarios de entrenamiento en control de corriente

B.1. Control de corriente monofásico con agente Q-learning

En este apéndice se muestran elementos, figuras, tablas y resúmenes adicionales a lo entregado en el documento, en lo que respecta al control de corriente monofásico.

B.1.1. Evaluación de mejores y peores casos de agentes según RMSE

Tabla B.1: Resultados del entrenamiento DQN en 1000 episodios.

Mejores						**Peores**					
Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	RMSE	Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	RMSE
071	32	1.e-4	0.97	256	10.35	088	64	1.e-3	0.95	256	86.33
086	64	1.e-3	0.97	128	10.47	085	64	1.e-3	0.95	128	77.99
062	32	5.e-4	0.97	256	10.50	003	24	5.e-3	0.99	64	73.79
018	24	1.e-3	0.99	256	10.58	028	24	1.e-4	0.95	64	68.93
072	32	1.e-4	0.99	256	10.69	078	64	5.e-3	0.99	128	68.19
025	24	5.e-4	0.95	256	10.72	001	24	5.e-3	0.95	64	65.09
075	64	5.e-3	0.99	64	10.84	092	64	5.e-4	0.97	64	50.71
106	64	1.e-4	0.95	256	10.89	077	64	5.e-3	0.97	128	48.98
043	32	5.e-3	0.95	256	10.92	069	32	1.e-4	0.99	128	48.55
024	24	5.e-4	0.99	128	10.97	068	32	1.e-4	0.97	128	47.01

Tabla B.2: Resultados del entrenamiento DQN según RMSE con distintas funciones de recompensas R.

Mejores							**Peores**						
Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	R	RMSE	Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	R	RMSE
062	32	5.e-4	0.95	256	2	9.02	072	32	5.e-4	0.99	256	4	43.84
092	32	5.e-3	0.97	256	4	9.02	023	24	5.e-4	0.99	256	3	43.75
027	24	1.e-3	0.95	256	3	9.03	034	24	1.e-3	0.99	256	2	43.71
090	32	5.e-3	0.97	256	2	9.04	039	24	5.e-3	0.95	256	3	43.68
004	24	1.e-4	0.95	256	4	9.04	031	24	1.e-3	0.97	256	3	43.54
026	24	1.e-3	0.95	256	2	9.05	056	32	1.e-4	0.97	256	4	43.53
028	24	1.e-3	0.95	256	4	9.05	052	32	1.e-4	0.95	256	4	43.49
003	24	1.e-4	0.95	256	3	9.06	079	32	1.e-3	0.97	256	3	43.47
006	24	1.e-4	0.97	256	2	9.06	050	32	1.e-4	0.95	256	2	43.47
066	32	5.e-4	0.97	256	2	9.06	094	32	5.e-3	0.99	256	2	43.44

Tabla B.3: Resultados del entrenamiento DQN según RMSE con distintos valores de TSF.

Mejores							**Peores**						
Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	TSF	RMSE	Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	TSF	RMSE
094	32	1e-3	0.97	256	5.e-4	9.08	055	24	1.e-4	0.95	128	1e-3	52.12
084	32	5e-3	0.99	256	5.e-4	9.68	012	24	1.e-3	0.95	256	1e-4	49.53
013	24	5e-3	0.97	256	1.e-3	9.87	030	24	1.e-4	0.95	256	1.e-4	49.15
108	32	5.e-4	0.99	256	5.e-4	9.99	027	24	5.e-4	0.99	256	1.e-4	48.13
099	32	5.e-4	0.97	128	1.e-3	10.00	009	24	5.e-3	0.99	256	1.e-4	46.58
093	32	1e-3	0.97	256	1.e-3	10.07	064	24	1.e-4	0.95	256	1.e-3	45.10
120	32	1e-4	0.99	256	5.e-4	10.11	090	32	1.e-3	0.99	128	5e-4	44.26
087	32	1e-3	0.97	128	1.e-3	10.14	035	24	1.e-3	0.99	256	5e-4	44.02
080	32	5e-3	0.95	256	5.e-4	10.31	006	24	5.e-3	0.97	256	1.e-4	43.58
018	24	1e-3	0.99	256	1.e-4	10.44	068	24	1.e-4	0.97	256	5e-4	42.63

Tabla B.4: Resultados entrenamiento final.

Mejores							**Peores**						
Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	$\varepsilon_{\text{decay}}$	RMSE	Run	N_n	α	γ	$\lvert\mathcal{M}\rvert$	$\varepsilon_{\text{decay}}$	RMSE
021	32	5.e-4	0.99	256	1.5e-5	8.90	007	24	5.e-4	0.95	256	1.5e-5	32.30
009	24	5.e-4	0.99	256	1.5e-5	8.95	015	32	5.e-3	0.99	256	1.5e-5	27.89
014	32	5.e-3	0.97	256	1.5e-5	8.96	019	32	5.e-4	0.95	256	1.5e-5	20.26
006	24	1.e-3	0.99	256	1.5e-5	8.96	003	24	5.e-3	0.99	256	1.5e-5	17.03
017	32	1.e-3	0.97	256	1.5e-5	8.98	004	24	1.e-3	0.95	256	1.5e-5	16.78
024	32	1.e-4	0.99	256	1.5e-5	9.00	013	32	5.e-3	0.95	256	1.5e-5	16.48
010	24	1.e-4	0.95	256	1.5e-5	9.01	005	24	1.e-3	0.97	256	1.5e-5	15.99
008	24	5.e-4	0.97	256	1.5e-5	9.05	018	32	1.e-3	0.99	256	1.5e-5	10.22
001	24	5.e-3	0.95	256	1.5e-5	9.05	002	24	5.e-3	0.97	256	1.5e-5	9.38
016	32	1.e-3	0.95	256	1.5e-5	9.07	020	32	5.e-4	0.97	256	1.5e-5	9.27

Apéndice C

Fundamentos matemáticos y límites físicos

C.1. Diferencias finitas

El método de diferencias finitas es una técnica numérica que permite aproximar derivadas mediante valores discretos de la función evaluados en puntos separados por un paso h. De este modo, una ecuación diferencial continua puede transformarse en una relación algebraica que aproxima el comportamiento dinámico del sistema.

La aproximación utilizada en este trabajo para la derivada primera de una función $x(t)$ es la llamada diferencia finita *forward*:

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) \approx \frac{x(t+h) - x(t)}{h} \tag{C.1}$$

C.2. Límite de corriente de inversor monofásico con carga RL

Es importante considerar el límite físico impuesto por la impedancia de la carga RL, que crece con la frecuencia por el término inductivo. La impedancia total es $|Z(f)| = \sqrt{R^2 + (2\pi f L)^2}$, de modo que con un inversor de dos niveles ideal la corriente máxima sinusoidal que puede sostenerse en régimen estacionario queda acotada por

$$I_{\mathrm{máx}}(f) \approx \frac{V_{dc}}{|Z(f)|}. \tag{C.2}$$

Para el caso de este estudio ($R = 5\,\Omega$, $L = 10\,\mathrm{mH}$, $V_{dc} = 200\,\mathrm{V}$), la amplitud máxima de corriente a 50 Hz es $I_{\mathrm{máx}}(50) \approx 33{,}9\,\mathrm{A}$. A partir de esta relación también se obtiene la frecuencia límite para una amplitud de referencia fija:

$$f_{A_{\mathrm{ref}}} = \frac{1}{2\pi L}\sqrt{\left(\frac{V_{dc}}{A_{\mathrm{ref}}}\right)^2 - R^2} \;\approx\; 308\ \mathrm{Hz}. \tag{C.3}$$

Apéndice D

Medidas estadísticas representativas en experimentos de aprendizaje por refuerzo

En el contexto del análisis del desempeño de agentes entrenados mediante métodos de aprendizaje por refuerzo, es habitual repetir múltiples ejecuciones independientes de un mismo experimento, variando parámetros entre procesos de entrenamiento. Debido a la naturaleza estocástica del aprendizaje y a la sensibilidad del agente a las condiciones iniciales, estas ejecuciones suelen presentar un grado significativo de variabilidad. Por ejemplo, ciertos episodios pueden conducir a una convergencia temprana y estable de la política, mientras que otros pueden exhibir divergencias transitorias o comportamientos subóptimos asociados a exploración excesiva o a la estructura de la función de recompensa.

Para caracterizar de manera adecuada el rendimiento global del agente frente a esta variabilidad, resulta necesario emplear medidas estadísticas que describan no sólo el valor central de una métrica de interés, como el error cuadrático medio (RMSE), sino también la dispersión o estabilidad de dicha medida entre ejecuciones. A continuación, se presentan las principales estadísticas utilizadas y su formulación matemática.

Media aritmética

La media aritmética del RMSE, calculada a partir de N ejecuciones independientes, se define como

$$\overline{\mathrm{RMSE}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{RMSE}_i, \tag{D.1}$$

donde RMSE_i corresponde al valor del error cuadrático medio obtenido en la ejecución i.

D.1. Mediana

La mediana se define como el percentil 50 de la distribución de valores de RMSE:

$$\mathrm{Mediana}(\mathrm{RMSE}) = \mathrm{percentil}_{50}(\{\mathrm{RMSE}_1, \mathrm{RMSE}_2, \ldots, \mathrm{RMSE}_N\}) \,. \tag{D.2}$$

A diferencia de la media, la mediana es robusta frente a valores atípicos y describe de forma más fidedigna el comportamiento típico del agente en la mayoría de las ejecuciones.

Desviación estándar

La dispersión de los valores de RMSE en torno a la media puede cuantificarse mediante la desviación estándar:

$$\sigma_{\mathrm{RMSE}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\mathrm{RMSE}_i - \overline{\mathrm{RMSE}}\right)^2}, \tag{D.3}$$

la cual permite evaluar la estabilidad del proceso de aprendizaje. Una desviación estándar baja indica que las distintas ejecuciones producen políticas consistentes, mientras que un valor elevado sugiere un comportamiento fuertemente dependiente de la inestabilidades estructurales en la configuración del agente.

Recomendación práctica

En análisis univariados de barridos de hiperparámetros, donde cada configuración se evalúa a partir de múltiples ejecuciones independientes, es recomendable reportar al menos tres estadísticas: la media, la mediana y la desviación estándar. La media ofrece una estimación del desempeño esperado, la mediana refleja el comportamiento típico del agente y la desviación estándar cuantifica su estabilidad. Este enfoque permite interpretar de manera más robusta el efecto de cada hiperparámetro y evita conclusiones sesgadas por ejecuciones atípicas [25].

Apéndice E

Motor de inducción trifásico

E.1. Parámetros de motor

Tabla E.1: Parámetros eléctricos y mecánicos de la máquina de inducción trifásica.

Parámetro	**Símbolo**	**Valor**	**Unidad**
Potencia nominal	P_{nom}	3 (2.24)	hp (kW)
Tensión nominal	V_{nom}	220	V
Número de pares de polos	p	2	–
Resistencia del estator	R_s	0.435	Ω
Resistencia del rotor (referida)	R'_r	0.816	Ω
Inductancia magnetizante	L_m	69.31	mH
Inductancia de dispersión del estator	L_{ls}	2.00	mH
Inductancia de dispersión del rotor (referida)	L'_{lr}	2.00	mH
Inductancia total del estator	L_s	71.31	mH
Inductancia total del rotor	L_r	71.31	mH
Inercia mecánica del rotor	J	0.089	kg·m^2

Tabla E.2: Valores nominales de la máquina y corrientes en el marco dq.

Magnitud	**Símbolo**	**Valor**	**Unidad**
Potencia nominal	P_{nom}	2.24	kW
Tensión nominal	V_{nom}	220	V
Velocidad nominal	n_{nom}	1500	rpm
Velocidad nominal	ω_{nom}	157.1	rad/s
Torque nominal	T_{nom}	11.2	N·m
Corriente nominal	I_{nom}	8.6	A
Corriente nominal peak	$I_{\text{nom,peak}}$	12.2	A
Componente d nominal	$i_{d,\text{nom}}$	5.0	A
Componente q nominal	$i_{q,\text{nom}}$	11.1	A

E.2. Parámetros de controladores

En esta sección se listan los parámetros de los reguladores utilizados en el capítulo de control de velocidad, así como los parámetros del lazo interno de corriente del FOC tradicional. Se deja el siguiente espacio para incorporar tablas de sintonía y valores finales.

E.2.1. PI del lazo de velocidad

Tabla E.3: Parámetros del PI del lazo de velocidad.

Parámetro	Símbolo	Valor
Ganancia proporcional	$K_{p,\omega}$	1.2
Ganancia integral	$K_{i,\omega}$	15.0

E.2.2. FOC tradicional: lazo interno de corriente

Tabla E.4: Parámetros del lazo interno de corriente del FOC tradicional.

Parámetro	Símbolo	Valor
Ganancia proporcional (d)	K_{p,i_d}	20
Ganancia integral (d)	K_{i,i_d}	2000
Ganancia proporcional (q)	K_{p,i_q}	20
Ganancia integral (q)	K_{i,i_q}	2000

Apéndice F

Resultados adicionales de control de velocidad

En la Figura F.1 se presenta el intento de arranque sin carga mecánica de la máquina asincrónica hacia la velocidad nominal utilizando el esquema QL–FOC. Durante los primeros instantes el lazo externo incrementa progresivamente la referencia de velocidad y, en consecuencia, el control demanda corriente de par para acelerar la máquina, observándose un crecimiento sostenido de ω.

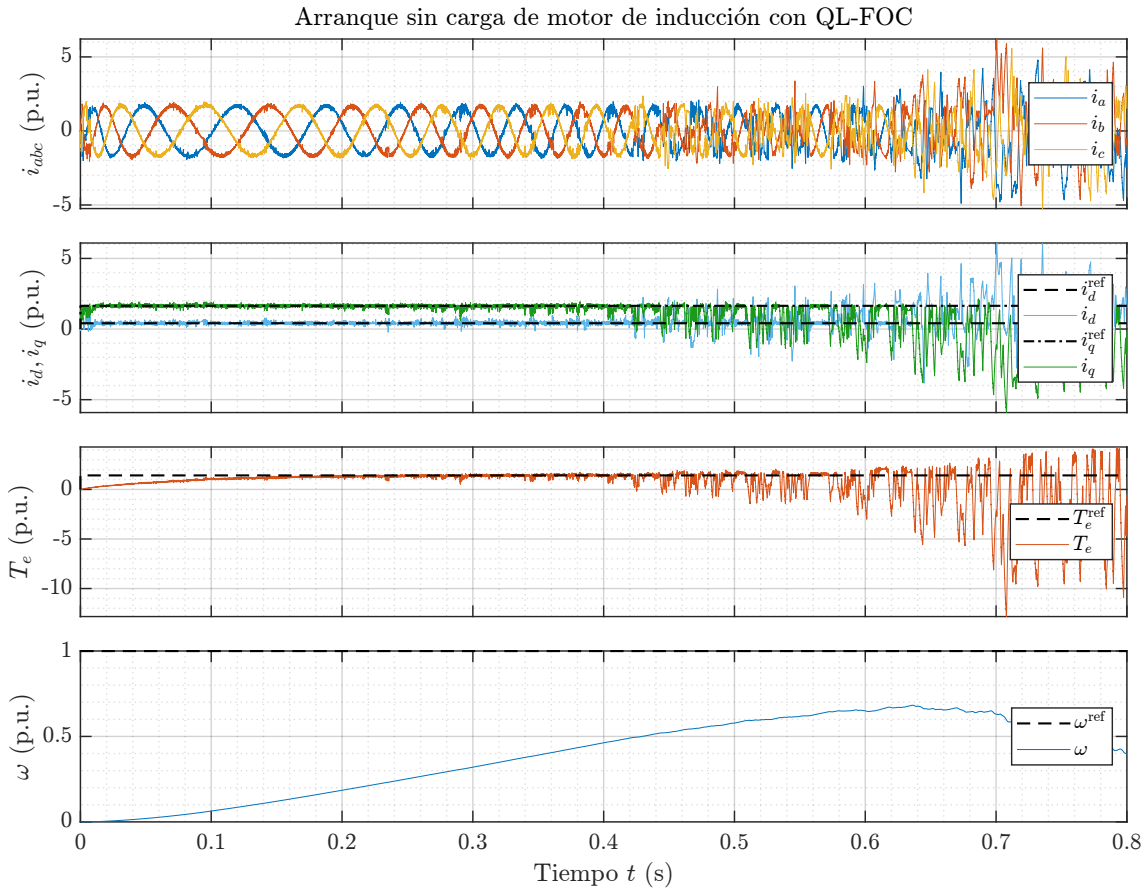


Figura F.1: Arranque de MAS a velocidad de referencia $\omega_{\text{ref}} = \omega_{\text{nom}}$ con QL-FOC.

Sin embargo, a partir de aproximadamente $t \approx 0{,}4$ s el comportamiento del sistema comienza a degradarse: las corrientes trifásicas i_{abc} aumentan su rizado y muestran oscilaciones de mayor amplitud, lo que se refleja también en el marco dq como un incremento de las fluctuaciones en i_d e i_q respecto de sus respectivas referencias. En paralelo, el torque eléctrico T_e presenta perturbaciones crecientes y variaciones abruptas, evidenciando que el control de corriente deja de sostener un seguimiento consistente bajo la demanda asociada al régimen de mayor velocidad.

Como consecuencia, el seguimiento de velocidad se vuelve inestable en el tramo final del experimento y la máquina no alcanza ni mantiene ω_{nom} de manera robusta. Este resultado respalda la decisión de realizar las pruebas del esquema QL–FOC a velocidad reducida, donde el controlador mantiene operación estable, mientras que el arranque a velocidad nominal conduce a una pérdida de control asociada a la incapacidad del agente para mantener de forma sostenida la componente i_q requerida durante la aceleración.

La Figura F.2 muestra el intento de arranque sin carga hacia la velocidad nominal incorporando una estrategia de flujo debilitado en el esquema QL–FOC. El objetivo de esta modificación es reducir la demanda de tensión del inversor a medida que aumenta la velocidad, permitiendo operar por sobre la zona base sin saturación del modulador.

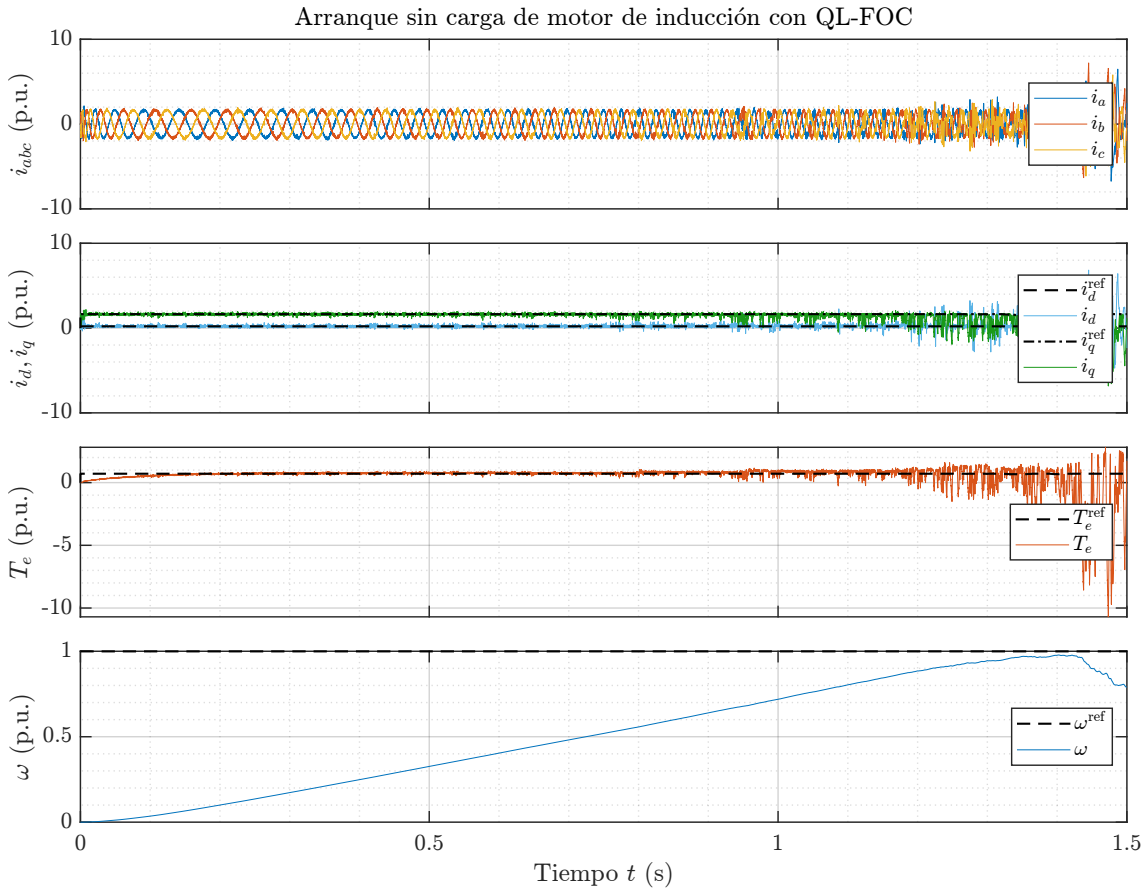


Figura F.2: Arranque de MAS a velocidad de referencia $\omega_{\text{ref}} = \omega_{\text{nom}}$ y flujo debilitado con QL-FOC.

Se observa que, durante gran parte del intervalo, las corrientes i_{abc} mantienen una amplitud acotada y un comportamiento más regular que en el caso sin debilitamiento, mientras que en el marco dq las componentes i_d e i_q permanecen cercanas a sus referencias, con rizado moderado. Esto sugiere que la reducción de flujo efectivamente alivia la exigencia del lazo interno de corriente

durante la aceleración.

No obstante, al aproximarse a la velocidad nominal (zona final del registro), reaparecen oscilaciones apreciables tanto en las corrientes como en el torque eléctrico T_e, junto con un deterioro del seguimiento de i_q. En consecuencia, aunque el flujo debilitado extiende el rango de operación estable y retrasa la pérdida de control, el esquema QL–FOC aún no logra mantener un arranque completamente robusto a ω_{nom}, evidenciando que la limitación principal se asocia a la capacidad del agente para sostener la componente de par requerida en condiciones de alta velocidad y dinámica más exigente.

Referencias

[1] Woorod A. Shutnan y Turki Y. Abdalla. «Artificial immune system based optimal fractional order PID control scheme for path tracking of robot manipulator». En: *2018 International Conference on Advance of Sustainable Engineering and its Application (ICASEA)*. 2018, págs. 19-24. DOI: 10.1109/ICASEA.2018.8370949.

[2] Shaswat Chirantan y Bibhuti Bhusan Pati. «MPC And HB-PWM Based Current Control Techniques for Three Phase Inverter Fed RL Load». En: *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)*. 2020, págs. 1-6. DOI: 10.1109/iSSSC50941.2020.9358858.

[3] D. G. Holmes y T. A. Lipo. *Pulse Width Modulation for Power Converters: Principles and Practice*. Capítulos 6–8: Control de corriente por histéresis y control deadbeat. Piscataway, NJ: IEEE Press, 2003. DOI: 10.1109/9780470546284.

[4] M. P. Kazmierkowski y L. Malesani. «Current control techniques for three-phase voltage-source PWM converters: A survey». En: *IEEE Transactions on Industrial Electronics* 45.5 (1998), págs. 691-703. DOI: 10.1109/41.720325.

[5] P. Kofinas, A. Dounis y G. Papadakis. «Adaptive fuzzy Q-Learning for control of distributed energy resources and electric vehicles». En: *Applied Energy* 219 (2018), págs. 486-499. DOI: 10.1016/j.apenergy.2018.03.039.

[6] T. Nguyen, H. Lee y J. Kim. «Deep reinforcement learning for power converter control: A review and case study». En: *IEEE Access* 8 (2020), págs. 185461-185473. DOI: 10.1109/ACCESS.2020.3029638.

[7] Yuchen Liu, Kian H. Low y Wenbo Shi. «Deep Reinforcement Learning for Power System Applications: An Overview». En: *Electric Power Systems Research* 190 (2021), pág. 106790. DOI: 10.1016/j.epsr.2020.106790.

[8] Ned Mohan, Tore M. Undeland y William P. Robbins. *Power Electronics: Converters, Applications, and Design*. 4th. Wiley, 2022. DOI: 10.1002/9781119700241.

[9] Bimal K. Bose. *Power Electronics and Motor Drives: Advances and Trends*. Academic Press, 2007. ISBN: 9780120884056.

[10] Paul C. Krause et al. *Analysis of Electric Machinery and Drive Systems*. 3rd. IEEE Press, 2013. DOI: 10.1002/9781118524336.

[11] Péter Vas. *Sensorless Vector and Direct Torque Control*. Oxford University Press, 1998. ISBN: 9780198564652.

[12] Richard S. Sutton y Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd. MIT Press, 2018. URL: http://incompleteideas.net/book/the-book-2nd.html.

[13] Richard Bellman y Stuart E. Dreyfus. *Applied Dynamic Programming*. Princeton University Press, 1962.

[14] C Watkins y P Dayan. «Technical Note Q-Learning. Machine Learning». En: 8(3) (1992), págs. 279-292. DOI: 10.1023/A:1022676722315.

[15] V. Ramesh et al. «Comparative analysis of Q-learning, SARSA, and deep Q-network for microgrid energy management». En: *Sustainability* 17.2 (2025), pág. 550. DOI: 10.3390/su17020550. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11698738/.

[16] Kurt Hornik, Maxwell Stinchcombe y Halbert White. «Multilayer feedforward networks are universal approximators». En: *Neural Networks* 2.5 (1989), págs. 359-366.

[17] Ian Goodfellow, Yoshua Bengio y Aaron Courville. *Deep Learning*. MIT Press, 2016.

[18] Volodymyr Mnih et al. «Playing Atari with Deep Reinforcement Learning». En: *arXiv preprint arXiv:1312.5602* (2013). DOI: 10.48550/arXiv.1312.5602. URL: https://arxiv.org/abs/1312.5602.

[19] J.N. Tsitsiklis y B. Van Roy. «An analysis of temporal-difference learning with function approximation». En: *IEEE Transactions on Automatic Control* 42.5 (1997), págs. 674-690. DOI: 10.1109/9.580874.

[20] Volodymyr Mnih et al. «Human-level control through deep reinforcement learning». En: *Nature* 518.7540 (2015), págs. 529-533.

[21] Vincent François-Lavet et al. «An Introduction to Deep Reinforcement Learning». En: *Foundations and Trends in Machine Learning* 11.3-4 (2018), págs. 219-354. DOI: 10.1561/2200000071.

[22] MathWorks. *Reinforcement Learning Toolbox Documentation*. Available at: https://www.mathworks.com/help/reinforcement-learning/. MathWorks. 2024.

[23] Hado Van Hasselt, Arthur Guez y David Silver. «Deep Reinforcement Learning with Double Q-learning». En: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[24] Timothy P. et al. Lillicrap. «Continuous control with deep reinforcement learning». En: *arXiv preprint arXiv:1509.02971* (2015).

[25] Peter Henderson et al. «Deep Reinforcement Learning that Matters». En: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018.

[26] Marlos C. Machado et al. «Revisiting the Arcade Learning Environment: Evaluation protocols and open problems for general agents». En: *JAIR*. Vol. 61. 2018, págs. 523-562.

[27] Yujia Zheng, Chao Liu y Xiaoguang Wang. «Reinforcement learning based control for nonlinear systems with action delay». En: *Neurocomputing* 333 (2019), págs. 1-9. DOI: 10.1016/j.neucom.2019.01.002.