

ACTAS

DE LAS

XXXVIII Jornadas de Automática

Gijón · Palacio de Congresos · 6, 7 y 8 de Septiembre de 2017



Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo



CEA
Comité Español
de Automática

Colabora

Gijón

Convention Bureau

Actas de

XXXVIII

Jornadas de Automática

© 2017 Universidad de Oviedo
© Los autores

Servicio de Publicaciones de la Universidad de Oviedo
Campus de Humanidades. Edificio de Servicios. 33011 Oviedo (Asturias)
Tel. 985 10 95 03 Fax 985 10 95 07
[http: www.uniovi.es/publicaciones](http://www.uniovi.es/publicaciones)
servipub@uniovi.es

DL AS 2749-2017

ISBN: 978-84-16664-74-0

Todos los derechos reservados. De conformidad con lo dispuesto en la legislación vigente, podrán ser castigados con penas de multa y privación de libertad quienes reproduzcan o plagien, en todo o en parte, una obra literaria, artística o científica, fijada en cualquier tipo y soporte, sin la preceptiva autorización.

Prefacio

Las *Jornadas de Automática* se celebran desde hace **40 años** en una universidad nacional facilitando el encuentro entre expertos en esta área en un foro que permite la puesta en común de las nuevas ideas y proyectos en desarrollo. Al mismo tiempo, propician la siempre necesaria colaboración entre investigadores del ámbito de la Ingeniería de Control y Automática, así como de campos afines, a la hora de abordar complejos proyectos de investigación multidisciplinares.

En esta ocasión, las Jornadas estarán organizadas por la Universidad de Oviedo y se han celebrado del 6 al 8 de septiembre de 2017 en el Palacio de Congresos de Gijón, colaborando tanto la Escuela Politécnica de Ingeniería de Gijón (EPI) como el Departamento de Ingeniería Eléctrica, Electrónica de Computadores y de Sistemas del que depende el Área de Ingeniería de Sistemas y Automática.

Además de las habituales actividades científicas y culturales, esta edición es muy especial al celebrarse el **50 aniversario de la creación de CEA**, Comité Español de Automática. Igualmente este año se conmemora el 60 aniversario de la Federación Internacional del Control Automático de la que depende CEA. Así se ha llevado a cabo la presentación del libro que se ha realizado bajo la coordinación de D. Sebastián Dormido, sobre la historia de la Automática en España en una sesión en la que han participado todos los ex-presidentes de CEA conjuntamente con el actual, D. Joseba Quevedo.

Igualmente hemos contado con la presencia de conferenciantes de prestigio para las sesiones plenarias, comunicaciones y ponencias orales en las reuniones de los 9 grupos temáticos, contribuciones en formato póster. Se ha celebrado también el concurso de CEABOT, así como una nueva Competición de Drones, con el ánimo de involucrar a más estudiantes de últimos cursos de Grado/Máster.

En el marco de las actividades culturales programadas se ha podido efectuar un recorrido en el casco antiguo situado en torno al Cerro de Santa Catalina y visitar la Laboral.

Gijón, septiembre de 2017

Hilario López
Presidente del Comité Organizador

Program Committee

Antonio Agudo	Institut de Robòtica i Informàtica Industrial
Rosa M Aguilar	University of La Laguna.
Luciano Alonso	University of Cantabria
Ignacio Álvarez García	Universidad de Oviedo
Antonio Javier Artuñedo García	Centre for Automation and Robotics (CSIC-UPM)
José M. Azorín	Miguel Hernandez University of Elche
Pedro Balaguer	Universitat Jaume I
Antonio Javier Barragán Piña	Universidad de Huelva
Alfonso Baños	Universidad de Murcia
Guillermo Bejarano	University of Seville
Gerardo Beruvides	Centro de Automática y Robótica
Carlos Bordons	University of Seville
Jose Manuel Bravo	University of Huelva
Jose Luis Calvo-Rolle	University of A Coruña
Fernando Castaño Romero	Centro de Automática y Robótica (UPM -CSIC)
José Luis Casteleiro-Roca	University of Coruña
Alvaro Castro-Gonzalez	Universidad Carlos III de Madrid
Ramon Costa-Castelló	Universitat Politècnica de Catalunya
Abel A. Cuadrado	University of Oviedo
Arturo De La Escalera	Universidad Carlos III de Madrid
Emma Delgado	Universidad de Vigo
Jose-Luis Diez	Universitat Politecnica de Valencia
Manuel Domínguez	Universidad de León
Juan Manuel Escaño	Universidad de Sevilla
Mario Francisco	University of Salamanca
Maria Jesus Fuente	Universidad de Valladolid
Juan Garrido	Universtiy of Cordoba
Antonio Giménez	Universidad de Almeria
Evelio Gonzalez	Universidad de La Laguna
José-Luis Guzmán	Universidad de Almería
Rodolfo Haber	Center for Automation and Robotics (UPM-CSIC)
César Ernesto Hernández	Universidad de Almería
Eloy Irigoyen	UPV/EHU
Agustin Jimenez	Universidad PolitÁcnica de Madrid
Emilio Jiménez	University of La Rioja
Jesus Lozano	Universidad de Extremadura
Jorge Luis Madrid	Centro de Automática y Robótica
Luis Magdalena	Universidad Politécnic de Madrid
David Martin Gomez	Universidad Carlos III de Madrid
Fernando Matia	Universidad Politecnica de Madrid
Joaquim Melendez	Universitat de Girona
Juan Mendez	Universidad de La Laguna
Luis Moreno	Universidad Carlos III de Madrid
María Dolores Moreno Rabel	Universidad de Extremadura
David Muñoz	Universidad de Sevilla
Antonio José Muñoz-Ramirez	Universidad de Málaga
Jose Luis Navarro	Universidad Politecnica de Valencia
Manuel G. Ortega	University of Seville
Andrzej Pawlowski	UNED
Mercedes Perez de La Parte	University of La Rioja
Ignacio Peñarrocha	Universitat Jaume I de Castelló, Spain
José Luis Pitarch	Universidad de Valladolid

Daniel Pérez	University of Oviedo
Emilio Pérez	Universitat Jaume I
Juan Pérez Oria	Universidad de Cantabria
MiguelÁngel Ridao	Universidad de Sevilla
Gregorio Sainz-Palmero	Universidad de Valladolid
Antonio Sala	Universitat Politecnica de Valencia
Ester Sales-Setién	Universitat Jaume I
Jose Sanchez	UNED
Javier Sanchis Saez	Universitat Politecnica de Valencia (UPV)
José Pedro Santos	ITEFI-CSIC
Matilde Santos	Universidad Complutense de Madrid
Alvaro Serna	University of Valladolid
José Enrique Simó	Universidad Politécnica de Valencia
José A. Somolinos	ETS I Navales. Universidad Politecnica de Madrid
Fernando Tadeo	Univ. of Valladolid
Alejandro Tapia	Universidad de Loyola Andalucía
David Tena	Universitat Jaume I
Jesús Torres	Universidad de La Laguna
Pedro M. Vallejo	Universidad de Salamanca
Guilherme Vianna	Universidad de Sevilla
Alejandro Vignoni	AI2 - UPV
Ramón Vilanova	UAB
Francisco Vázquez	Universidad de Cordoba
Jesús M. Zamarreño	University of Valladolid

Revisores Adicionales

Al-Kaff, Abdulla

Balbastre, Patricia
Beltrán de La Cita, Jorge
Bermudez-Cameo, Jesus
Blanco-Claraco, Jose-Luis
Blanes, Francisco
Bonin-Font, Francisco

Cancela, Brais

Ferraz, Luis

Garita, Cesar
Gimenez, Antonio
Gruber, Patrick
Guindel, Carlos

Hernandez Ruiz, Alejandro
Hernandez, Daniel

Jardón Huete, Alberto

López, Amable

Marin, Raul
Marín Plaza, Pablo
Mañanas, Miguel Angel
Morales, Rafael
Moreno, Francisco-Angel

Nuñez, Luis Ramón

Ponz Vila, Aurelio
Posadas-Yague, Juan-Luis
Poza-Luján, Jose-Luis
Pumarola, Albert

Raya, Rafael
Revestido Herrero, Elías
Rocon, Eduardo
Ruiz Sarmiento, José Raúl
Ruiz, Adria

Torres, Jose Luis

Vaquero, Victor

Table of Contents

Ingeniería de Control	
TÚNEL DE AGUA PARA PRUEBAS Y CARACTERIZACIÓN DE DISEÑOS EXPERIMENTALES DE TURBINAS HIDROCINÉTICAS	1
<i>Eduardo Alvarez, Manuel Rico-Secades, Antonio Javier Calleja Rodríguez, Joaquín Fernández Francos, Aitor Fernández Jiménez, Mario Alvarez Fernández and Samuel Camba Fernández</i>	
Reduction of population variability in protein expression: A control engineering approach.	8
<i>Yadira Boada, Alejandro Vignoni and Jesús Picó</i>	
CONTROL ROBUSTO DEL PH EN FOTOBIORREACTORES MEDIANTE RECHAZO ACTIVO DE PERTURBACIONES	16
<i>José Carreño, Jose Luis Guzman, José Carlos Moreno and Rodolfo Villamizar</i>	
Control reset para maniobra de cambio de carril y validación con CarSim.....	23
<i>Miguel Cerdeira, Pablo Falcón, Antonio Barreiro, Emma Delgado and Miguel Díaz-Cacho</i>	
Maniobra de aterrizaje automática de una Cessna 172P modelada en FlightGear y controlada desde un programa en C	31
<i>Mario de La Rosa, Antonio Javier Gallego and Eduardo Fernández</i>	
Alternativas para el control de la red eléctrica aislada en parques eólicos marinos	38
<i>Carlos Díaz-Sanahuja, Ignacio Peñarrocha, Ricardo Vidal-Albalade and Ester Sales-Setién</i>	
CONTROL PREDICTIVO DISTRIBUIDO UTILIZANDO MODELOS DIFUSOS PARA LA NEGOCIACIÓN ENTRE AGENTES	46
<i>Lucía Fargallo, Silvana Roxani Revollar Chavez, Mario Francisco, Pastora Vega and Antonio Cembellín</i>	
Control Predictivo en el espacio de estados de un captador solar tipo Fresnel	54
<i>Antonio Javier Gallego, Mario de La Rosa and Eduardo Fernández</i>	
Control predictivo para la operación eficiente de una planta formada por un sistema de desalación solar y un invernadero.....	62
<i>Juan Diego Gil Vergel, Lidia Roca, Manuel Berenguel, Alba Ruiz Aguirre, Guillermo Zaragoza and Antonio Giménez</i>	
Depuración de Aguas Residuales en la Industria 4.0	70
<i>Jesus Manuel Gomez-De-Gabriel, Ana María Jiménez Arévalo, Laura Eiroa Mateo and Fco. Javier Fernández-De-Cañete-Rodríguez</i>	
Control robusto con QFT del pH en un fotobioreactor raceway	77
<i>Ángeles Hoyo Sánchez, Jose Luis Guzman, Jose Carlos Moreno and Manuel Berenguel</i>	
Revisión sistemática de la literatura en ingeniería de sistemas. Caso práctico: técnicas de estimación distribuida de sistemas ciberfísicos	84
<i>Carmelina Ierardi, Luis Orihuela Espina, Isabel Jurado Flores, Álvaro Rodríguez Del Nozal and Alejandro Tapia Córdoba</i>	
Desarrollo de un Controlador Predictivo para Autómatas programables basado en la normativa IEC 61131-3.....	92
<i>Pablo Krupa, Daniel Limon and Teodoro Alamo</i>	
Diseño de un emulador de aerogenerador de velocidad variable DFIG y control de pitch...100	
<i>Manuel Lara Ortiz, Juan Garrido Jurado and Francisco Vázquez Serrano</i>	

Observación de la fracción de agua líquida en pilas de combustible tipo PEM de cátodo abierto.....	108
<i>Julio Luna and Ramon Costa-Castelló</i>	
Control Predictivo Basado en Datos.....	115
<i>José María Manzano, Daniel Limón, Teodoro Álamo and Jan Peter Calliess</i>	
Control MPC basado en un modelo LTV para seguimiento de trayectoria con estabilidad garantizada.....	122
<i>Sara Mata, Asier Zubizarreta, Ione Nieva, Itziar Cabanes and Charles Pinto</i>	
Implementación y evaluación de controladores basados en eventos en la norma IEC-61499.	130
<i>Oscar Miguel-Escrig, Julio-Ariel Romero-Pérez and Esteban Querol-Dolz</i>	
AUTOMATIZACIÓN Y MONITORIZACIÓN DE UNA INSTALACIÓN DE ENSAYO DE MOTORES.....	138
<i>Alfonso Poncela Méndez, Miguel Ochoa Vega, Eduardo J. Moya de La Torre and F. Javier García Ruíz</i>	
OPTIMIZACIÓN Y CONTROL EN CASCADA DE TEMPERATURA DE RECINTO MEDIANTE SISTEMAS DE REFRIGERACIÓN.....	146
<i>David Rodríguez, José Enrique Alonso Alfaya, Guillermo Bejarano Pellicer and Manuel G. Ortega</i>	
Diseño LQ e implementación distribuida para la estimación de estado.....	154
<i>Álvaro Rodríguez Del Nozal, Luis Orihuela, Pablo Millán Gata, Carmelina Ierardi and Alejandro Tapia Córdoba</i>	
Estimación de fugas en un sistema industrial real mediante modelado por señales aditivas.	160
<i>Ester Sales-Setién, Ignacio Peñarrocha and David Tena</i>	
Advanced control based on MPC ideas for offshore hydrogen production.....	167
<i>Alvaro Serna, Fernando Tadeo and Julio. E Normey-Rico</i>	
Transfer function parameters estimation by symmetric send-on-delta sampling.....	174
<i>José Sánchez, María Guinaldo, Sebastián Dormido and Antonio Visioli</i>	
An Estimation Approach for Process Control based on Asymmetric Oscillations.....	181
<i>José Sánchez, María Guinaldo Losada, Sebastian Dormido, José Luis Fernández Marrón and Antonio Visioli</i>	
Robust PI controller for disturbance attenuation and its application for voltage regulation in islanded microgrid.....	189
<i>Ramon Vilanova, Carles Pedret and Orlando Arrieta</i>	
Infraestructura para explotación de datos de un simulador azucarero.....	197
<i>Jesús M. Zamarréño, Cristian Pablos, Alejandro Merino, L. Felipe Acebes and De Prada César</i>	
<hr/> Automar <hr/>	
INFRAESTRUCTURA PARA ESTUDIAR ADAPTABILIDAD Y TRANSPARENCIA EN EL CENTRO DE CONTROL VERSÁTIL.....	203
<i>Juan Antonio Bonache Seco, José Antonio Lopez Orozco, Eva Besada Portas and Jesús Manuel de La Cruz</i>	
ARQUITECTURA DE CONTROL HÍBRIDA PARA LA NAVEGACIÓN DE VEHÍCULOS SUBMARINOS NO TRIPULADOS.....	211
<i>Francisco J. Lastra, Jesús A. Trujillo, Francisco J. Velasco and Elías Revestido</i>	

Exploración y Reconstrucción 3D de Fondos Marinos Mediante AUVs y Sensores Acústicos	218
<i>Oscar L. Manrique Garcia, Mario Andrei Garzon Oviedo and Antonio Barrientos</i>	
AUTOMATIZACIÓN DE MANIOBRAS PARA UN TEC DE 2GdL	226
<i>Marina Pérez de La Portilla, José Andrés Somolinos Sánchez, Amable López Piñeiro, Rafael Morales Herrera and Eva Segura</i>	
MERBOTS PROJECT: OVERALL DESCRIPTION, MULTISENSORY AUTONOMOUS PERCEPTION AND GRASPING FOR UNDERWATER ROBOTICS INTERVENTIONS	232
<i>Pedro J. Sanz, Raul Marin, Antonio Peñalver, David Fornas and Diego Centelles</i>	
<hr/> Bioingeniería <hr/>	
MARCADORES CUADRADOS Y DEFORMACIÓN DE OBJETOS EN NAVEGACIÓN QUIRÚRGICA CON REALIDAD AUMENTADA	238
<i>Eliana Aguilar, Oscar Andres Vivas and Jose Maria Sabater-Navarro</i>	
Entrenamiento robótico de la marcha en pacientes con Parálisis Cerebral: definición de objetivos, propuesta de tratamiento e implementación clínica preliminar	244
<i>Cristina Bayón, Teresa Martín-Lorenzo, Beatriz Moral-Saiz, Óscar Ramírez, Álvaro Pérez-Somarriba, Sergio Lerma-Lara, Ignacio Martínez and Eduardo Rocon</i>	
PREDICCIÓN DE ACTIVIDADES DE LA VIDA DIARIA EN ENTORNOS INTELIGENTES PARA PERSONAS CON MOVILIDAD REDUCIDA	251
<i>Arturo Bertomeu-Motos, Santiago Ezquerro, Juan Antonio Barios, Luis Daniel Lledó, Francisco Javier Badesa and Nicolas Garcia-Aracil</i>	
Sistema de Visión Estereoscópico para el guiado de un Robot Quirúrgico en Operaciones de Cirugía Laparoscópica HALS.....	256
<i>Carlos Castedo Hernández, Rafael Estop Remacha, Eusebio de La Fuente López and Lidia Santos Del Blanco</i>	
Head movement assessment of cerebral palsy users with severe motor disorders when they control a computer thought eye movements.....	264
<i>Alejandro Clemotte, Miguel A. Velasco and Eduardo Rocon</i>	
Diseño de un sensor óptico de fuerza para exoesqueletos de mano.....	270
<i>Jorge Diez Pomares, Andrea Blanco Ivorra, José María Catalan Orts, Francisco Javier Badesa Clemente, José María Sabater and Nicolas Garcia Aracil</i>	
POSIBILIDADES DEL USO DE TRAMAS ARTIFICIALES DE IMAGEN MOTORA PARA UN BCI BASADO EN EEG	276
<i>Josep Dinarès-Ferran, Christoph Guger and Jordi Solé-Casals</i>	
EFFECTOS SOBRE LA ERD EN TAREAS DE CONTROL DE EXOESQUELETO DE MANO EMPLEANDO BCI.....	282
<i>Santiago Ezquerro, Juan Antonio Barios, Arturo Bertomeu-Motos, Luisa Lorente, Nuria Requena, Irene Delegido, Francisco Javier Badesa and Nicolas Garcia-Aracil</i>	
Formulación Topológica Adaptada para la Simulación y Control de Exoesqueletos Accionados con Transmisiones Harmonic Drive.....	288
<i>Andres Hidalgo Romero and Eduardo Rocon</i>	

Identificación de contracciones isométricas de la extremidad superior en pacientes con lesión medular incompleta mediante características espectrales de la electromiografía de alta densidad (HD-EMG)	296
<i>Mislav Jordanic, Mónica Rojas-Martínez, Joan Francesc Alonso, Carolina Migliorelli and Miguel Ángel Mañanas</i>	
Diseño de una plataforma para analizar el efecto de la estimulación mecánica aferente en el temblor de pacientes con temblor esencial	302
<i>Julio S. Lora, Roberto López, Jesús González de La Aleja and Eduardo Rocon</i>	
DEFINICIÓN DE UN PROTOCOLO PARA LA MEDIDA PRECISA DEL RANGO CERVICAL EMPLEANDO TECNOLOGÍA INERCIAL	308
<i>Álvaro Martín, Rafael Raya, Cristina Sánchez, Rodrigo Garcia-Carmona, Oscar Ramirez and Abraham Otero</i>	
SISTEMA BRAIN-COMPUTER INTEFACE DE NAVEGACIÓN WEB ORIENTADO A PERSONAS CON GRAVE DISCAPACIDAD.....	313
<i>Víctor Martínez-Cagigal, Javier Gómez-Pilar, Daniel Álvarez, Eduardo Santamaría-Vázquez and Roberto Hornero</i>	
ESTRATEGIAS DE NEUROESTIMULACIÓN TRANSCRANEAL POR CORRIENTE DIRECTA PARA MEJORA COGNITIVA	320
<i>Silvia Moreno Serrano, Mario Ortiz and José María Azorín Poveda</i>	
COMPARATIVA DE ALGORITMOS PARA LA DETECCIÓN ONLINE DE IMAGINACIÓN MOTORA DE LA MARCHA BASADO EN SEÑALES DE EEG	328
<i>Marisol Rodríguez-Ugarte, Irma Nayeli Angulo Sherman, Eduardo Iáñez and Jose M. Azorin</i>	
DETECCIÓN, MEDIANTE UN GUANTE SENSORIZADO, DE MOVIMIENTOS SELECCIONADOS EN UN SISTEMA ROBOTIZADO COLABORATIVO PARA HALS	334
<i>Lidia Santos, José Luis González, Eusebio de La Fuente, Juan Carlos Fraile and Javier Pérez Turiel</i>	
BIOSENSORES PARA CONTROL Y SEGUIMIENTO PATOLOGÍAS REUMATOIDES	340
<i>Amparo Tirado, Raúl Marín, José V Martí, Miguel Belmonte and Pedro Sanz</i>	
Assessment of tremor severity in patients with essential tremor using smartwatches	347
<i>Miguel A. Velasco, Roberto López-Blanco, Juan P. Romero, M. Dolores Del Castillo, J. Ignacio Serrano, Julián Benito-León and Eduardo Rocon</i>	
INTERFAZ CEREBRO-ORDENADOR PARA EL CONTROL DE UNA SILLA DE RUEDAS A TRAVÉS DE DOS PARADIGMAS DE NAVEGACIÓN	353
<i>Fernández-Rodríguez Álvaro, Velasco-Álvarez Francisco and Ricardo Ron-Angevin</i>	
<hr/>	
Control Inteligente	
<hr/>	
Aprendizaje por Refuerzo para sistemas lineales discretos con dinámica desconocida: Simulación y Aplicación a un Sistema Electromecánico	360
<i>Henry Diaz, Antonio Sala and Leopoldo Armesto</i>	
Diseño de sistemas de control en cascada clásico y borroso para el seguimiento de trayectorias	368
<i>Javier G. Gonzalez, Rodolfo Haber, Fernando Matia and Marcelino Novo</i>	

ANÁLISIS FORMAL DE LA DINÁMICA DE SISTEMAS NO LINEALES MEDIANTE REDES NEURONALES.....	376
<i>Eloy Irigoyen, Mikel Larrea, A. Javier Barragán, Miguel Ángel Martínez and José Manuel Andújar</i>	
Predicción de la energía renovable proveniente del oleaje en las islas de Fuerteventura y Lanzarote.	384
<i>G.Nicolás Marichal, Deivis Avila, Ángela Hernández, Isidro Padrón and José Ángel Rodríguez</i>	
Aplicación de Redes Neuronales para la Estimación de la Resistencia al Avance en Buques	393
<i>Daniel Marón Blanco and Matilde Santos</i>	
Novel Fuzzy Torque Vectoring Controller for Electric Vehicles with per-wheel Motors.....	401
<i>Alberto Parra, Martín Dendaluze, Asier Zubizarreta and Joshué Pérez</i>	
REPOSTAJE EN TIERRA DE UN AVIÓN MEDIANTE ALGORITMOS GENÉTICOS .	408
<i>Elías Plaza and Matilde Santos</i>	
VISUALIZACIÓN WEB INTERACTIVA PARA EL ANÁLISIS DEL CHATTER EN LAMINACIÓN EN FRÍO.....	416
<i>Daniel Pérez López, Abel Alberto Cuadrado Vega and Ignacio Díaz Blanco</i>	
BANCADA PARA ANÁLISIS INTELIGENTE DE DATOS EN MONITORIZACIÓN DE SALUD ESTRUCTURAL.....	424
<i>Daniel Pérez López, Diego García Pérez, Ignacio Díaz Blanco and Abel Alberto Cuadrado Vega</i>	
CONTROL DE UN VEHÍCULO CUATRIRROTOR BASADO EN REDES NEURONALES.....	431
<i>Jesus Enrique Sierra and Matilde Santos</i>	
CONTROL PREDICTIVO FUZZY CON APLICACIÓN A LA DEPURACIÓN BIOLÓGICA DE FANGOS ACTIVADOS.....	437
<i>Pedro M. Vallejo Llamas and Pastora Vega Cruz</i>	
<hr/> Educación en Automática <hr/>	
REFLEXIONES SOBRE EL VALOR DOCENTE DE UNA COMPETICION DE DRONES EN LA EDUCACIÓN PARA EL CONTROL.....	445
<i>Ignacio Díaz Blanco, Alvaro Escanciano Urigüen, Antonio Robles Alvarez and Hilario López García</i>	
Uso del Haptic Paddle con aprendizaje basado en proyectos.....	451
<i>Juan M. Gandarias, Antonio José Muñoz-Ramírez and Jesus Manuel Gomez-De-Gabriel</i>	
REPRESENTACION INTEGRADA DE ACCIONAMIENTOS MECANICOS Y CONTROL DE EJES ORIENTADA A LA COMUNICACIÓN Y DOCENCIA EN MECATRONICA.....	457
<i>Julio Garrido Campos, David Santos Esterán, Juan Sáez López and José Ignacio Armesto Quiroga</i>	
Construcción y modelado de un prototipo fan & plate para prácticas de control automático	465
<i>Cristina Lampon, Javier Martin, Ramon Costa-Castelló and Muppaneni Lokesh Chowdary</i>	

EDUCACION EN AUTOMATICA E INDUSTRIA 4.0 MEDIANTE LA APLICACIÓN DE TECNOLOGÍAS 3D	471
<i>Jose Ramon Llata, Esther Gonzalez-Sarabia, Carlos Torre-Ferrero and Ramon Sancibrian</i>	
Desarrollo e implementación de un sistema de control en una planta piloto hibrida.....	479
<i>Maria P. Marcos, Cesar de Prada and Jose Luis Pitarch</i>	
LA INFORMÁTICA INDUSTRIAL EN LAS INGENIERÍAS INDUSTRIALES	486
<i>Rogelio Mazaeda, Eusebio de La Fuente López, José Luis González, Eduardo J. Moya de La Torre, Miguel Angel García Blanco, Javier García Ruiz, María Jesús de La Fuente Aparicio, Gregorio Sainz Palmero and Smaranda Cristea</i>	
Ventajas docentes de un flotador magnético para la experimentación de técnicas control ..	495
<i>Eduardo Montijano, Carlos Bernal, Carlos Sagües, Antonio Bono and Jesús Sergio Artal</i>	
PROGRAMACIÓN ATRACTIVA DE PLC	502
<i>Eduardo J. Moya de La Torre, F. Javier García Ruíz, Alfonso Poncela Méndez and Victor Barrio Lángara</i>	
MODERNIZACIÓN DE EQUIPO FEEDBACK MS-150 PARA EL APRENDIZAJE ACTIVO EN INGENIERÍA DE CONTROL	510
<i>Perfecto Reguera Acevedo, Miguel Ángel Prada Medrano, Antonio Morán Álvarez, Juan José Fuertes Martínez, Manuel Domínguez González and Serafín Alonso Castro</i>	
INNOVACIÓN PEDAGÓGICA EN LA FORMACIÓN DEL PERFIL PROFESIONAL PARA EL DESARROLLO DE PROYECTOS DE AUTOMATIZACIÓN INDUSTRIAL A TRAVÉS DE UNA APROXIMACIÓN HOLÍSTICA.....	517
<i>Juan Carlos Ríos, Zaneta Babel, Daniel Martínez, José María Paredes, Luis Alonso, Pablo Hernández, Alejandro García, David Álvarez, Jorge Miranda, Constantino Manuel Valdés and Jesús Alonso</i>	
Aprendiendo Simulación de Eventos Discretos con JaamSim	522
<i>Enrique Teruel and Rosario Aragüés</i>	
RED NEURONAL AUTORREGRESIVA NO LINEAL CON ENTRADAS EXÓGENAS PARA LA PREDICCIÓN DEL ELECTROENCEFALOGRAMA FETAL...	528
<i>Rosa M Aguilar, Jesús Torres and Carlos Martín</i>	
ANÁLISIS DEL COEFICIENTE DE TRANSFERENCIA DE MATERIA EN REACTORES RACEWAYS.....	534
<i>Marta Barceló, Jose Luis Guzman, Francisco Gabriel Acién, Ismael Martín and Jorge Antonio Sánchez</i>	
MODELADO DINÁMICO DE UN SISTEMA DE ALMACENAMIENTO DE FRÍO VINCULADO A UN CICLO DE REFRIGERACIÓN	539
<i>Guillermo Bejarano Pellicer, José Joaquín Suffo, Manuel Vargas and Manuel G. Ortega</i>	
Predictor Intervalar basado en hiperplano soporte	547
<i>José Manuel Bravo Caro, Manuel Vasallo Vázquez, Emilian Cojocarú and Teodoro Alamo Cantarero</i>	
Dynamic simulation applied to refinery hydrogen networks	555
<i>Anibal Galan Prado, Cesar De Prada, Gloria Gutierrez, Rafael Gonzalez and Daniel Sarabia</i>	

APROXIMACIÓN DE MODELOS ALGEBRAICOS MEDIANTE ALAMO Y ECOSIMPRO	563
<i>Carlos Gómez Palacín, José Luis Pitarch, Gloria Gutiérrez and Cesar De Prada</i>	
A Causal Model to Analyze Aircraft Collision Avoidance Deadlock Scenarios	569
<i>Miquel Àngel Piera Eroles, Julia de Homdedeu, Maria Del Mar Tous, Thimjo Koca and Marko Radanovic</i>	
ONLINE DECISION SUPPORT FOR AN EVAPORATION NETWORK	575
<i>José Luis Pitarch, Marc Kalliski, Carlos Gómez Palacín, Christian Jasch and Cesar De Prada</i>	
Predicción de la irradiancia a partir de datos de satélite mediante deep learning	582
<i>Javier Pérez, Jorge Segarra-Tamarit, Hector Beltran, Carlos Ariño, José Carlos Alfonso Gil, Aleks Attanasio and Emilio Pérez</i>	
MODELO DINÁMICO ORIENTADO AL TRATAMIENTO Y SEGUIMIENTO DE LA LEUCEMIA MIELOIDE CRÓNICA	589
<i>Gabriel Pérez Rodríguez and Fernando Morilla</i>	
Modelado y optimización de la operación de un sistema de bombeo de múltiples depósitos	596
<i>Roberto Sanchis Llopis and Ignacio Peñarrocha</i>	
DEVELOPMENT OF A GREY MODEL FOR A MEDIUM DENSITY FIBREBOARD DRYER IN ECOSIMPRO	604
<i>Pedro Santos, Jose Luis Pitarch and César de Prada</i>	
DETECCIÓN AUTOMÁTICA DE FALLOS MEDIANTE MONITORIZACIÓN Y OPTIMIZACIÓN DE LAS FECHAS DE LIMPIEZA PARA INSTALACIONES FOTOVOLTAICAS	611
<i>Jorge Segarra-Tamarit, Emilio Pérez, Hector Beltran, Enrique Belenguer and José Luis Gandía</i>	
Modelado de micro-central hidráulica para el diseño de controladores con aplicación en regiones aisladas de Honduras	618
<i>Alejandro Tapia Córdoba, Pablo Millán Gata, Fabio Gómez-Estern Aguilar, Carmelina Ierardi and Álvaro Rodríguez Del Nozal</i>	
FRAMEWORK PARA EL MODELADO DE UN LAGO DE DATOS	626
<i>J.M Torres, R.M. Aguilar, C.A. Martin and S. Diaz</i>	
SIMULADOR CARDIOVASCULAR PARA ENSAYO DE ROBOTS DE NAVEGACION AUTONOMA	633
<i>José Emilio Traver, Juan Francisco Ortega Morán, Ines Tejado, J. Blas Pagador, Fei Sun, Raquel Pérez-Aloe, Blas M. Vinagre and F. Miguel Sánchez Margallo</i>	
PLANIFICACION DE LA PRODUCCION BASADA EN CONTROL PREDICTIVO PARA PLANTAS TERMOSOLARES	641
<i>Manuel Jesús Vasallo Vázquez, José Manuel Bravo Caro, Emilian Cojocarú and Manuel Emilio Gegundez Arias</i>	
Evaluación multicriterio para la optimización de redes de energía	649
<i>Ascensión Zafra Cabeza, Rafael Espinosa, Miguel Àngel Ridao Carlini and Carlos Bordóns Alba</i>	
Percibiendo el entorno en los robots sociales del RoboticsLab	657
<i>Fernando Alonso Martín, Jose Carlos Castillo Montoya, Àlvaro Castro-Gonzalez, Juan José Gamboa, Marcos Maroto Gómez, Sara Marqués Villaroya, Antonio J. Pérez Vidal and Miguel Àngel Salichs</i>	

DISEÑO DE UNA PRÓTESIS DE MANO ADAPTABLE AL CRECIMIENTO	664
<i>Marta Ayats and Raul Suarez</i>	
COOPERATIVISMO BIOINSPIRADO BASADO EN EL COMPORTAMIENTO DE LAS HORMIGAS	672
<i>Brayan Bermudez, Kristel Novoa and Miguel Valbuena</i>	
PROCEDIMIENTO DE DISEÑO DE UN EXOESQUELETO DE MIEMBRO SUPERIOR PARA SOPORTE DE CARGAS	680
<i>Andrea Blanco Ivorra, Jorge Diez Pomares, David Lopez Perez, Francisco Javier Badesa Clemente, Miguel Ignacio Sanchez and Nicolas Garcia Aracil</i>	
Estructura de control en ROS y modos de marcha basados en máquinas de estados de un robot hexápodo	686
<i>Raúl Cebolla Arroyo, Jorge De Leon Rivas and Antonio Barrientos</i>	
USING AN UAV TO GUIDE THE TELEOPERATION OF A MOBILE MANIPULATOR	694
<i>Josep Arnau Claret and Luis Basañez</i>	
Estudio de los patrones de marcha para un robot hexápodo en tareas de búsqueda y rescate	701
<i>Jorge De León Rivas and Antonio Barrientos</i>	
SISTEMA DE INTERACCIÓN VISUAL PARA UN ROBOT SOCIAL	709
<i>Mario Domínguez López, Eduardo Zalama Casanova, Jaime Gómez GarcÍa-Bermejo and Samuel Marcos Pablos</i>	
Mejora del Comportamiento Proxémico de un Robot Autónomo mediante Motores de Inteligencia Artificial Desarrollados para Plataformas de Videojuegos	717
<i>David Fernández Chaves, Javier Monroy and Javier Gonzalez-Jimenez</i>	
Micrófonos de contacto: una alternativa para sensado táctil en robots sociales	724
<i>Juan José Gamboa, Fernando Alonso Martín, Jose Carlos Castillo, Marcos Maroto Gómez and Miguel A. Salichs</i>	
Clasificación de información táctil para la detección de personas	732
<i>Juan M. Gandarias, Jesús M. Gómez-De-Gabriel and Alfonso García-Cerezo</i>	
Planificación para interceptación de objetivos: Integración del Método Fast Marching y Risk-RRT	738
<i>David Alfredo Garzon Ramos, Mario Andrei Garzon Oviedo and Antonio Barrientos</i>	
ESTABILIZACIÓN DE UNA BOLA SOBRE UN PLANO UTILIZANDO UN ROBOT PARALELO 6-RSS	746
<i>Daniel González, Lluís Ros and Federico Thomas</i>	
TELEOPERACIÓN DE INSTRUMENTOS QUIRÚRGICOS ARTICULADOS	754
<i>Ana Gómez Delgado, Carlos Perez-Del-Pulgar, Antonio Reina Terol and Victor Muñoz Martinez</i>	
CONTROL OF A ROBOTIC ARM FOR TRANSPORTING OBJECTS BASED ON NEURO-FUZZY LEARNING VISUAL INFORMATION	760
<i>Juan Hernández Vicén, Santiago Martínez de La Casa Díaz and Carlos Balaguer</i>	
PLATAFORMA BASADA EN LA INTEGRACIÓN DE MATLAB Y ROS PARA LA DOCENCIA DE ROBÓTICA DE SERVICIO	766
<i>Carlos G. Juan, Jose Maria Vicente, Alvaro Garcia and Jose Maria Sabater-Navarro</i>	

Estimadores de fuerza y movimiento para el control de un robot de rehabilitación de extremidad superior.....	772
<i>Aitziber Mancisidor, Asier Zubizarreta, Itziar Cabanes, Pablo Bengoa and Asier Brull</i>	
Definiendo los elementos que constituyen un robot social portable de bajo coste	780
<i>Marcos Maroto Gómez, José Carlos Castillo, Fernando Alonso-Martín, Juan José Gamboa, Sara Marqués Villarroya and Miguel Ángel Salichs</i>	
Interfaces táctiles para Interacción Humano-Robot	787
<i>Sara Marqués Villarroya, Jose Carlos Castillo Montoya, Fernando Alonso Martín, Marcos Maroto Gómez, Juan José Gamboa and Miguel A. Salichs</i>	
HERRAMIENTAS DE ENTRENAMIENTO Y MONITORIZACIÓN PARA EL DESMINADO HUMANITARIO	793
<i>Hector Montes, Roemi Fernandez, Pablo Gonzalez de Santos and Manuel Armada</i>	
Control a Baja Velocidad de una Rueda con Motor de Accionamiento Directo mediante Ingeniería Basada en Modelos	799
<i>Antonio José Muñoz-Ramírez, Jesús Manuel Luque-Bedmar, Jesus Manuel Gomez-De-Gabriel, Anthony Mandow, Javier Serón and Alfonso Garcia-Cerezo</i>	
SIMULACIÓN DE VEHÍCULOS AUTÓNOMOS USANDO V-REP BAJO ROS	806
<i>Cándido Otero Moreira, Enrique Paz Domonte, Rafael Sanz Dominguez, Joaquín López Fernández, Rafael Barea, Eduardo Romera, Eduardo Molinos, Roberto Arroyo, Luís Miguel Bergasa and Elena López</i>	
Cinemática y prototipado de un manipulador paralelo con centro de rotación remoto para robótica quirúrgica.....	814
<i>Francisco Pastor, Juan M. Gandarias and Jesús M. Gómez-De-Gabriel</i>	
ANÁLISIS DE ESTABILIDAD DE SINGULARIDADES AISLADAS EN ROBOTS PARALELOS MEDIANTE DESARROLLOS DE TAYLOR DE SEGUNDO ORDEN.....	821
<i>Adrián Peidro Vidal, Óscar Reinoso, Arturo Gil, José María Marín and Luis Payá</i>	
INTERFAZ DE CONTROL PARA UN ROBOT MANIPULADOR MEDIANTE REALIDAD VIRTUAL	829
<i>Elena Peña-Tapia, Juan Jesús Roldán, Mario Garzón, Andrés Martín-Barrio and Antonio Barrientos</i>	
Evolución de la robótica social y nuevas tendencias.....	836
<i>Antonio J. Pérez Vidal, Alvaro Castro-Gonzalez, Fernando Alonso Martín, Jose Carlos Castillo Montoya and Miguel A. Salichs</i>	
DISEÑO MECÁNICO DE UN ASISTENTE ROBÓTICO CAMARÓGRAFO CON APRENDIZAJE COGNITIVO	844
<i>Irene Rivas-Blanco, M Carmen López-Casado, Carlos Pérez-Del-Pulgar, Francisco García-Vacas, Víctor Fernando Muñoz, Enrique Bauzano and Juan Carlos Fraile</i>	
CÁLCULO DE FUERZAS DE CONTACTO PARA PRENSIONES BIMANUALES.....	852
<i>Francisco Abiud Rojas-De-Silva and Raul Suarez</i>	
Modelado del Contexto Geométrico para el Reconocimiento de Objetos.....	860
<i>José Raúl Ruiz Sarmiento, Cipriano Galindo and Javier Gonzalez-Jimenez</i>	
Estimación Probabilística de Áreas de Emisión de Gases con un Robot Móvil Mediante la Integración Temporal de Observaciones de Gas y Viento	868
<i>Carlos Sanchez-Garrido, Javier Monroy and Javier Gonzalez-Jimenez</i>	

MANIPULADOR AÉREO CON BRAZOS ANTROPOMÓRFICOS DE ARTICULACIONES FLEXIBLES	876
<i>Alejandro Suarez, Guillermo Heredia and Anibal Ollero</i>	
EVALUACIÓN DE UN ENTORNO DE TELEOPERACIÓN CON ROS	864
<i>David Vargas Frutos, Juan Carlos Ramos Martínez, José Luis Samper Escudero, Miguel Ángel Sánchez-Urán González and Manuel Ferre Pérez</i>	
<hr/> Sistemas de Tiempo Real <hr/>	
GENERACIÓN DE CÓDIGO IEC 61131-3 A PARTIR DE DISEÑOS EN GRAFCET....	892
<i>María Luz Alvarez Gutierrez, Isabel Sarachaga Gonzalez, Arantzazu Burgos Fernandez, Nagore Iriondo Urbistazu and Marga Marcos Muñoz</i>	
CONTROL EN TIEMPO REAL Y SUPERVISIÓN DE PROCESOS MEDIANTE SERVIDORES OPC-UA	900
<i>Francisco Blanes Noguera and Andrés Benlloch Faus</i>	
Control de la Ejecución en Sistemas de Criticidad Mixta	906
<i>Alfons Crespo, Patricia Balbastre, Jose Simo and Javier Coronel</i>	
GENERACIÓN AUTOMÁTICA DEL PROYECTO DE AUTOMATIZACIÓN TIA PORTAL PARA MÁQUINAS MODULARES	913
<i>Darío Orive, Aintzane Armentia, Eneko Fernandez and Marga Marcos</i>	
DDS en el desarrollo de sistemas distribuidos heterogéneos con soporte para criticidad mixta	921
<i>Hector Perez and J. Javier Gutiérrez</i>	
ARQUITECTURA DISTRIBUIDA PARA EL CONTROL AUTÓNOMO DE DRONES EN INTERIOR	929
<i>Jose-Luis Poza-Luján, Juan-Luis Posadas-Yaguë, Giovanni-Javier Tipantuña-Topanta, Francisco Abad and Ramón Mollá</i>	
Ingeniería Conducida por Modelos en Sistemas de Automatización Flexibles	935
<i>Rafael Priego, Elisabet Estévez, Dario Orive, Isabel Sarachaga and Marga Marcos</i>	
Estudio e implementación de Middleware para aplicaciones de control distribuido	942
<i>Jose Simo, Jose-Luis Poza-Lujan, Juan-Luis Posadas-Yaguë and Francisco Blanes</i>	
<hr/> Visión por Computador <hr/>	
Real-Time Image Mosaicking for Mapping and Exploration Purposes	948
<i>Abdulla Al-Kaff, Juan Camilo Soto Triviño, Raúl Sosa San Frutos, Arturo de La Escalera and José María Armingol Moreno</i>	
ALGORITMO DE SLAM UTILIZANDO APARIENCIA GLOBAL DE IMÁGENES OMNIDIRECCIONALES	956
<i>Yerai Berenguer, Luis Payá, Mónica Ballesta, Luis Miguel Jiménez, Sergio Cebollada and Oscar Reinoso</i>	
Medición de Oximetría de Pulso mediante Imagen fotopletismográfica.....	964
<i>Juan-Carlos Cobos-Torres, Jordan Ortega Rodríguez, Pablo J. Alhama Blanco and Mohamed Abderrahim</i>	
Algoritmo de captura de movimiento basado en visión por computador para la teleoperación de robots humanoides.....	970
<i>Juan Miguel Garcia Haro and Santiago Martinez de La Casa</i>	

COMPARACIÓN DE MÉTODOS DE DETECCIÓN DE ROSTROS EN IMÁGENES DIGITALES	976
<i>Natalia García Del Prado, Victor Gonzalez Castro, Enrique Alegre and Eduardo Fidalgo Fernández</i>	
LOCALIZACIÓN DEL PUNTO DE FUGA PARA SISTEMA DE DETECCIÓN DE LÍNEAS DE CARRIL	983
<i>Manuel Ibarra-Arenado, Tardi Tjahjadi, Sandra Robla-Gómez and Juan Pérez-Oria</i>	
Oculus-Crawl, a Software Tool for Building Datasets for Computer Vision Tasks	991
<i>Iván De Paz Centeno, Eduardo Fidalgo Fernández, Enrique Alegre Gutiérrez and Wesam Al Nabki</i>	
Clasificación automática de obstáculos empleando escáner láser y visión por computador ..	999
<i>Aurelio Ponz, Fernando Garcia, David Martin, Arturo de La Escalera and Jose Maria Armingol</i>	
T-SCAN: OBTENCIÓN DE NUBES DE PUNTOS CON COLOR Y TEMPERATURA EN INTERIOR DE EDIFICIOS	1007
<i>Tomás Prado, Blanca Quintana, Samuel A. Prieto and Antonio Adan</i>	
EVALUACIÓN DE MÉTODOS PARA REALIZAR RESÚMENES AUTOMÁTICOS DE VÍDEOS	1015
<i>Pablo Rubio, Eduardo Fidalgo, Enrique Alegre and Víctor González</i>	
SIMULADOR PARA LA CREACIÓN DE MUNDOS VIRTUALES PARA LA ASISTENCIA A PERSONAS CON MOVILIDAD REDUCIDA EN SILLA DE RUEDAS ..	1023
<i>Carlos Sánchez Sánchez, María Cidoncha Jiménez, Emiliano Pérez, Ines Tejado and Blas M. Vinagre</i>	
Calibración Extrínseca de un Conjunto de Cámaras RGB-D sobre un Robot Móvil	1031
<i>David Zúñiga-Nöel, Rubén Gómez Ojeda, Francisco-Ángel Moreno and Javier González Jiménez</i>	

Aprendizaje por Refuerzo para sistemas lineales discretos con dinámica desconocida: Simulación y Aplicación a un Sistema Electromecánico

Henry Díaz, Leopoldo Armesto, Antonio Sala
 {hendia@posgrado larmesto@idf asala@isa}.upv.es
 Universitat Politècnica de València,
 C/Camino de Vera s/n, 46022, Valencia, España

Resumen

El aprendizaje por refuerzo es una técnica que se utiliza en la búsqueda de soluciones en sistemas de decisión secuencial. Una gran parte de los algoritmos usados en el aprendizaje por refuerzo se fundamentan en la programación dinámica, se considera que el aprendizaje por refuerzo es una extensión de la programación dinámica que proporciona soluciones sin la necesidad de conocer el modelo de comportamiento del sistema. Estas técnicas combinan algunas características del control óptimo y control adaptativo para el diseño de controladores realimentados. Se describen los algoritmos básicos del aprendizaje por refuerzo para la implementación de soluciones en sistemas discretos deterministas. Finalmente, se realizaron pruebas prácticas de la implementación del algoritmo de aprendizaje *Q-Learning* en un péndulo de un grado de libertad, con el objetivo de verificar si el algoritmo de aprendizaje converge y proporciona un controlador estabilizante.

Palabras clave Aprendizaje por refuerzo, *Q-Learning*, control óptimo, control adaptativo óptimo, programación dinámica .

1. Introducción

El aprendizaje por refuerzo (RL, *reinforcement learning*) es un conjunto de técnicas para resolver problemas de decisión secuenciales, en los cuales las decisiones son aplicadas al sistema con el objetivo de obtener una respuesta deseada[12][16]. Este tipo de problemas secuenciales aparecen en una amplia variedad de campos entre los que podemos mencionar el control automático, inteligencia artificial, robótica, control de procesos, entre otras[10]. El aprendizaje por refuerzo a diferencia de las técnicas de programación dinámica (DP, *dynamic programming*) no requiere del conocimiento del modelo[2][3]. Debido a los múltiples orígenes del RL es común encontrar en la literatura los mismos conceptos con diferentes definiciones, por ejemplo: programación neuro-dinámica[4], programación dinámica aproximada[10], programación dinámica adaptativa[9] entre otros.

Los principales elementos y su interacción del pro-



Figura 1: Elementos de la DP y RL y su flujo de interacción[5].

blema a resolver en la DP y RL son representados en la Figura 1. Se muestra la manera en la que un agente interactúa con el entorno mediante tres señales: una señal de estado del entorno, una señal de acción que permite al agente influenciar el estado del entorno y una señal escalar de recompensa, la cual proporciona al agente información sobre la calidad de la acción que acaba de realizar en el estado actual. En cada instante temporal, el agente recibe una medida del estado y realiza una acción. Como consecuencia de la acción realizada se produce una transición del entorno a un nuevo estado. Además se genera una señal de recompensa que evalúa la calidad de dicha transición. Entonces el agente recibe el nuevo estado y el ciclo completo se repite [12].

El agente selecciona la acción realizada en cada estado de acuerdo a una *política*. La política es una función que mapea los estados a acciones. El objetivo del agente es aprender una política que maximice la cantidad total de recompensa recibida, es decir, la recompensa acumulada a largo plazo[12][4][5].

Considerando las características que presentan los controladores proporcionados por las técnicas de RL se analizan e implementan algoritmos para sistemas discretos lineales de los cuales se desconoce su comportamiento dinámico[6][7].

El paper está estructurado de la siguiente forma: en la sección 2 se describen los conceptos básicos de la programación dinámica y el aprendizaje

por refuerzo. La sección 3 describe el aprendizaje por refuerzo para sistemas lineales discretos. En la sección 4 se realiza la simulación del aprendizaje en sistemas lineales. En la sección 5 se describe la aplicación del algoritmo de aprendizaje *Q-Learning* a un péndulo de un grado de libertad y se finaliza con la sección 6 de conclusiones.

2. Programación dinámica y aprendizaje por refuerzo

La programación dinámica es una parte fundamental de la teoría de control óptimo. En un problema de control óptimo, el objetivo es desarrollar un controlador que minimice una medida del comportamiento de un sistema dinámico a lo largo del tiempo [4], esta medida de comportamiento es evaluada con un índice de costo o función de valor y este índice o valor puede ser definido en términos de objetivos de optimalidad [8].

El RL tiene sus orígenes en el campo de la inteligencia artificial y se encuentra inspirado en los mecanismos de aprendizaje biológico. Específicamente, tiene sus raíces en el condicionamiento operante entre las diferentes formas que un individuo puede responder ante una misma situación, aquellas que estén acompañadas de una satisfacción (refuerzo positivo), estarán más firmemente conectadas a dicha situación de repetirse.

Los algoritmos DP para encontrar una política óptima requieren de un modelo MDP *Markov decision processes* incluyendo la dinámica de transición y la función de refuerzo [4], en general muchos problemas de decisión toman como marco de referencia los MDP incluidos los sistemas de control realimentados [7], un estudio detallado y amplio sobre MDP se puede consultar en [11].

Los algoritmos de RL son libres de modelo [12][4] lo que les hace muy útiles cuando la obtención del modelo de un proceso es demasiado dificultosa o muy costosa de ser implementada. Estos algoritmos usan datos obtenidos del proceso, estos datos pueden ser un conjunto de trayectorias, una simple trayectoria o un conjunto de muestras, lo que implica trabajar con un número limitado de datos que provienen proceso. Mientras los algoritmos de DP pueden usar el modelo para obtener cualquier número de muestras de transición de cualquier par estado-acción.

2.1. Criterios de optimalidad

El objetivo de los algoritmos de DP/RL es encontrar una política que maximice la recompensa obtenida por el agente a lo largo del tiempo. Escoger entre los criterios de optimalidad está relacionado

con el problema del aprendizaje. La mayor parte de los algoritmos de DP y RL emplean el criterio de optimalidad de horizonte infinito descontado (ecuación 1) debido a que posee propiedades teóricas que lo hacen más adecuado para el análisis matemático [4].

$$\sum_{t=0}^{\infty} \gamma^t r_t \quad (1)$$

Donde r_t es el refuerzo instantáneo y $\gamma \in [0, 1)$ es el factor de descuento.

2.2. Función de valor y ecuaciones de Bellman

Las funciones de valor son el punto de unión entre el sistema y el criterio de optimalidad. Una *función de valor* es una estimación de la bondad que supone para un agente estar en un determinado estado cuando se sigue una política fija. Existen dos tipos de funciones de valor: función V , que estima la bondad de estar en un estado, y función Q que estima la bondad de realizar una acción en un estado. Usando el modelo de horizonte infinito con descuento la función de valor puede ser expresada así:

$$V^\pi(x) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (2)$$

Una función de valor estado-acción similar:

$$Q^\pi(x, u) = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \quad (3)$$

Una de las características fundamentales de las funciones de valor es que satisfacen ciertas propiedades recursivas. Para cualquier política π y cualquier estado x la expresión en la ecuación 2 puede ser definida recursivamente en términos de la llamada *ecuación de Bellman* [2].

$$\begin{aligned} V^\pi(x) &= r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots \\ &= r_t + \gamma V^\pi(x_{t+1}) \end{aligned} \quad (4)$$

La meta buscada es encontrar la *mejor* política, por ejemplo la que reciba el mayor retorno. Esto significa maximizar la ecuación 2 para todos los estados $x \in X$. Una *política óptima*, denotada π^* , es tal que $V^{\pi^*}(x) \geq V^\pi(x)$ para todo $x \in X$ y todas las políticas π . Se puede demostrar que la solución óptima $V^* = V^{\pi^*}$ satisface las siguiente ecuación:

$$V^*(x) = \max_u [r(x, u) + \gamma V^*(x')] \quad (5)$$

Esta ecuación se denomina la *Ecuación de optimalidad de Bellman*. Y establece que el valor de un estado bajo una política óptima debe ser igual

al retorno esperado para la mejor acción en ese estado. Para seleccionar una acción óptima dada la función de valor óptima V^* se puede aplicar la siguiente regla:

$$\pi^*(x) = \operatorname{argm\acute{a}x}_u [r(x, u) + \gamma V^\pi(x')] \quad (6)$$

La denominación de esta política es *política voraz* (*greedy policy*), se denota $\pi_{\text{greedy}}(V)$. Esta política selecciona la mejor acción usando la función de valor V . Análogamente el valor óptimo estado-acción es:

$$Q^*(x, u) = r(x, u) + \gamma \operatorname{m\acute{a}x}_{u'} Q^*(x', u') \quad (7)$$

Las Q -funciones son muy útiles debido a que hacen innecesaria la suma ponderada sobre las diferentes alternativas usando la función de transición. Esa es la razón por la cual en el enfoque libre de modelo donde no se conoce la función de transición ni la función de recompensa son aprendidas en lugar de las V -funciones. La relación entre Q^* y V^* esta dada por:

$$V^*(x) = \operatorname{m\acute{a}x}_u Q^*(x, u) \quad (8)$$

La selección de la acción óptima esta dada por:

$$\pi^*(x) = \operatorname{argm\acute{a}x}_u Q^*(x, u) \quad (9)$$

Es decir, la mejor acción es la acción que tiene la mayor utilidad esperada sobre la base de posibles estados próximos resultantes de tomar esa acción. Los algoritmos de DP y RL de acuerdo a como se obtienen la política óptima se clasifican en algoritmos de Iteración de función de valor (VI, *Value Iteration*), buscan el valor óptimo de la función de valor, que consiste en el máximo refuerzo de cada estado o de cada par estado-acción. Algoritmos de Iteración de política (PI, *Policy Iteration*), evalúan las políticas a través de construir sus funciones de valor (en lugar de la función de valor óptima), y utilizan estas funciones de valor para hallar nuevas y mejores políticas.

Hay varios métodos para la implementación de los algoritmos VI y PI. Los tres principales son: cálculo exacto, métodos de Monte Carlo y aprendizaje por diferencias temporales (TD, *Temporal difference*) [4][12][10]. Los dos últimos métodos pueden ser implementados sin el conocimiento de la dinámica del sistema, el método de diferencias temporales es el que se toma como referencia en las secciones posteriores.

2.2.1. Temporal difference

Temporal difference hace referencia a una familia de métodos para estimar, o predecir, la función V de una política fija, aunque como veremos en secciones posteriores, el concepto del aprendizaje TD puede ser extendido al caso de funciones Q [12]. En

los métodos TD la función V se estima en base a otras estimaciones previas, técnica que recibe el nombre de *bootstrapping* [12]. Cada vez que el agente realiza una acción el algoritmo TD utiliza la recompensa generada y la estimación actual de V para realizar una nueva estimación de acuerdo a la expresión:

$$V_{k+1}(x_k) = V_k(x_k) + \alpha_k [r_{k+1} + \gamma V_k(x_{k+1}) - V_k(x_k)] \quad (10)$$

donde $\alpha_k \in [0, 1]$ es la secuencia de tasas de aprendizaje que determina la cantidad con la que se actualiza el valor del estado x_k . El término entre corchetes, se conoce como diferencia temporal y da nombre al método, es la diferencia entre la nueva estimación de la función V , $r_{k+1} + \gamma V_k(x_{k+1})$, y la estimación en el instante temporal anterior, $V_k(x_k)$.

3. Aprendizaje por refuerzo y control adaptativo óptimo para sistemas lineales en tiempo discreto

El análisis físico de los sistemas utilizando por ejemplo la mecánica lagrangiana o la mecánica hamiltoniana que son una reformulación de la mecánica clásica proporcionan descripciones de los sistemas en términos de ecuaciones diferenciales ordinarias no lineales. Discretizando obtenemos una representación de los sistemas en ecuaciones en diferencias [8].

Considerando un sistema discreto representado por las siguiente ecuación en diferencias

$$x_{k+1} = f(x_k) + g(x_k)u_k \quad (11)$$

donde el estado $x_k \in R^n$ y la acción de control $u_k \in R^m$. Una política de control esta definida como una función del espacio de estados al espacio de control $h(\cdot) : R^n \rightarrow R^m$. Lo que significa que por cada estado se define una acción de control dada por

$$u_k = h(x_k) \quad (12)$$

Una política es simplemente un controlador realimentado. A partir de definir una función costo se obtiene la función de valor [7][6][8].

$$V^h(x_k) = \sum_{i=k}^{\infty} \gamma^{i-k} (x_i^T H_x x_i + u_i^T H_u u_i) \quad (13)$$

con un factor de descuento $0 < \gamma \leq 1$, $H_x \in \mathbb{R}^{n_x \times n_x}$ y $H_u \in \mathbb{R}^{n_u \times n_u}$, son las matrices de pesos de la función de costo cuadrático y $u_k = h(x_k)$ una política de control realimentada. El costo de cada etapa

$$r(x_k, u_k) = x_k^T H_x x_k + u_k^T H_u u_k \quad (14)$$

es considerado como cuadrático en u_k para simplificar el desarrollo, pero puede ser cualquier función de control definida positiva. Se asume que el sistema es estabilizante en un conjunto $\Omega \in R^n$, lo que significa que existe una política de control $u_k = h(x_k)$ que el sistema en lazo cerrado $x_{k+1} = f(x_k) + g(x_k)h(x_k)$ es asintóticamente estable en Ω . Una política se dice que es *admisibile* si esta es estabilizante y proporciona un costo finito $V^h(x_k)$ para la trayectorias en Ω [8][1]. Para sistemas determinísticos discretos, el valor óptimo esta dado por la ecuación de optimalidad de Bellman

$$V^*(x_k) = \min_{h(\cdot)}(r(x_k, h(x_k)) + \gamma V^*(x_{k+1})) \quad (15)$$

Que es justamente la ecuación Hamilton-Jacobi-Bellman(HJB)[8] en tiempo discreto. Y tenemos que la política óptima es

$$h^*(x_k) = \operatorname{argmín}_{h(\cdot)}(r(x_k, h(x_k)) + \gamma V^*(x_{k+1})) \quad (16)$$

Para el regulador lineal cuadrático para sistemas discretos(DT LQR) tenemos,

$$x_{k+1} = Ax_k + Bu_k \quad (17)$$

$$V^h(x_k) = \frac{1}{2} \sum_{i=k}^{\infty} \gamma^{i-k} (x_i^T H_x x_i + u_i^T H_u u_i) \quad (18)$$

Notar que desde el punto de vista de los sistemas de control el objetivo que busca el aprendizaje por refuerzo es encontrar una política óptima que minimice el coste acumulado.

Iteración de política(PI) usando aprendizaje por diferencias temporales[7]

Inicialización

Seleccionar cualquier política de control admisible $h_0(x_k)$.

Hacer para $j = 0$ hasta converger

Evaluación de Política

$$V_{j+1}(x_k) = r(x_k, h_j(x_k)) + \gamma V_{j+1}(x_{k+1}) \quad (19)$$

Mejora de Política

$$h_{j+1}(x_k) = \operatorname{argmín}_{h(\cdot)}(r(x_k, h(x_k)) + \gamma V_{j+1}(x_{k+1})) \quad (20)$$

o

$$h_{j+1}(x_k) = -\frac{\gamma}{2} R^{-1} g^T(x_k) \nabla V_{j+1}(x_{k+1}) \quad (21)$$

donde $\nabla V(x) = \delta V(x)/\delta x$ es el gradiente de la función de valor, interpretado aquí como un vector columna. En el método de iteración de la función de valor se realiza de manera similar, pero el procedimiento de evaluación de la política se realiza de la siguiente forma.

Iteración de la función de valor(VI) usando aprendizaje por diferencias temporales[7]

Actualización de la función valor en cada paso

Actualización del valor usando

$$V_{j+1}(x_k) = r(x_k, h_j(x_k)) + \gamma V_j(x_{k+1}) \quad (22)$$

En VI se puede seleccionar cualquier política de control inicial $h_0(x_k)$, no necesariamente admisible o estabilizante.

3.1. Aproximación de la función de valor

Para implementaciones prácticas de PI y VI para sistemas dinámicos con infinitos espacios de estado y de acciones es aproximar la función de valor usando una estructura de un aproximador adecuado en términos de parámetros desconocidos[6]. Así, los parámetros desconocidos son ajustados en línea exactamente como en un sistema de identificación. Esta idea de la *aproximación de la función de valor* (VFA) fue usada por Werbos[15][14] y llamada programación dinámica aproximada(ADP) o programación dinámica adaptativa. Esta fue usada por Bertsekas y Tsitsiklis[4] y la llamó programación neurodinámica.

En el caso del LQR es conocido que el valor es cuadrático en el estado para alguna matriz *kernel* P [8].

$$V(x_k) = \frac{1}{2} x_k^T P x_k = \frac{1}{2} (\operatorname{vec}(P))^T (x_k \otimes x_k) \equiv \bar{p}^T \phi(x_k) \quad (23)$$

El producto de Kronecker \otimes permite escribir esta forma cuadrática como una lineal en vector de parámetros $\bar{p} = \operatorname{vec}(P)$, que se forma apilando la columnas de la matriz P [7]. El vector $\phi(x_k) = \bar{x}_k = x_k \otimes x_k$ es el vector polinomial cuadrático que contiene todos los posibles pares de productos de n componentes de x_k . Notar que P es simétrica y tiene solamente $n(n+1)/2$ elementos independientes, removiendo los términos redundantes en $x_k \otimes x_k$ para definir un conjunto de base cuadrática $\phi(x_k)$ con $n(n+1)/2$ elementos independientes[7][6].

Se asume que la ecuación de Bellman tiene una solución local suave.

$$V(x) = \sum_{i=1}^{\infty} w_i \varphi_i(x) \equiv W^t \phi(x) + \varepsilon_L(x) \quad (24)$$

donde el vector de base $\phi(x) = [\varphi_1(x) \ \varphi_2(x) \ \dots \ \varphi_L(x)] : R^n \rightarrow R^L$ y $\varepsilon_L(x)$ converge uniformemente a cero mientras el número de términos $L \rightarrow \infty$ [7][5][3].

3.2. Control adaptativo óptimo en sistemas lineales discretos con dinámica desconocida

El método de RL *Q-learning* proporciona un algoritmo de control adaptativo que converge en línea a la solución de control óptima para sistemas en los que se desconoce completamente su dinámica. Este método resuelve la ecuación de Bellman y las ecuación HJB en tiempo real a través de la medición de datos a lo largo de las trayectorias del sistema, sin conocer la dinámica $f(x_k), g(x_k)$ [7].

Q-learning[13] es un método simple de RL que trabaja para sistemas desconocidos, esto es, para sistemas los cuales se desconoce completamente su dinámica. *Q-learning* aprende la función Q usando el método de diferencias temporales(TD) y realizando una acción u_k y midiendo en cada etapa el resultado del conjunto de datos de experiencia (x_k, x_{k+1}, r_k) consistentemente en el estado actual, el estado próximo y el costo resultante[8].

El algoritmo *Q-Learning* puede ser fácilmente desarrollado para sistemas dinámicos discretos usando aproximaciones de la función Q, en [8] se desarrolla y se muestra las principales ecuaciones para el *Q-Learning* en sistemas discretos.

Tomando como referencia [7], tenemos que para un sistema no lineal la función Q es parametrizada como

$$Q(x, u) = W^T \phi(z)$$

para algún vector de parámetros desconocido W y un conjunto de vectores base $\phi(z)$. Para un DT LQR, $\phi(z)$ es un conjunto base cuadrático formado por componentes de estado y entrada. Por lo tanto, el error TD es

$$e_k = -W^T \phi(z_k) + r(x_k, u_k) + \gamma W^T \phi(z_{k+1}) \quad (25)$$

sobre el cual los algoritmos PI y VI pueden basarse. Considerando el algoritmo PI el paso de evaluación de una función Q es

$$W_{j+1}^T (\phi(z_k) - \gamma \phi(z_{k+1})) = r(x_k, h_j(x_k)) \quad (26)$$

y el paso de mejora de la política es

$$h_{j+1}(x_k) = \underset{u}{\operatorname{argmín}} (W_{j+1}^T \phi(x_k, u)), x \in X \quad (27)$$

Q-learning usando VI esta dado por

$$W_{j+1}^T \phi(z_k) = r(x_k, h_j(x_k)) - \gamma W_j^T \phi(z_{k+1}) \quad (28)$$

y la ecuación 27. Estas ecuaciones no requieren conocimiento de la dinámica $f(\cdot), g(\cdot)$.

Para implementaciones en línea, para resolver la ecuación 26 se puede usar LS por lotes o RLS para

el vector de parámetros W_{j+1} obteniendo el vector de regresión $\phi(z_k) - \gamma \phi(z_{k+1})$, o en 28 usando el vector de regresión $\phi(z_k)$. Los datos observados en cada instante de tiempo son $(z_k, z_{k+1}, r(x_k, u_k))$ con $z_k \equiv [x_k^T u_k^T]^T$, con $u_{k+1} = h_j(x_{k+1})$ y $h_j(x_k)$ la política actual. Se debe agregar ruido de exploración a la entrada de control para obtener una excitación persistente.

Después de la convergencia de los parámetros de la función Q, la actualización de la acción es realizada. Esto se realiza fácilmente sin conocer la dinámica del sistema debido a que la función Q contiene u_k como uno de sus argumentos así que $\partial(W_{j+1}^T \phi(x_k, u))/\partial u$ puede ser explícitamente calculada.

$$\frac{\partial Q(x, u)}{\partial u} = \left(\frac{\partial z}{\partial u}\right)^T \left(\frac{\partial \phi(z)}{\partial z}\right)^T$$

$$W = [0_{m,n} \quad I_m] \nabla \phi^T W$$

donde $0_{m,n} \in R^{m \times n}$ es una matriz de ceros. El vector de base $\phi(z) = z \otimes z \in R^{n+m^2}$ es el vector polinomial cuadrático que contiene todos los posibles pares de productos de los $n+m$ componentes de z . Se define $N = n+m$, entonces

$$\nabla \phi^T = \frac{\partial \phi^T}{\partial z} = (I_N \otimes z + z \otimes I_N)^T \in R^{N \times N^2} \quad (29)$$

4. Simulaciones implementadas

Las simulaciones realizadas en esta sección consideran sistemas discretos lineales en los cuales la dinámica del sistema no es conocida.

4.1. Aprendizaje por refuerzo para un Sistema lineal discreto de segundo orden usando PI (*Policy Iteration*).

En esta simulación se muestra el uso del algoritmo PI para resolver la DT ARE sin conocer la dinámica del sistema, es decir, para el caso del LQR se desconoce las matrices A y B del sistema. Las matrices solo se usan para generar las trayectorias y adquirir los datos que el algoritmo requiere. Por lo tanto, el modelo del sistema a considerar aquí es $x_{k+1} = Ax_k + Bu_k$, donde

$$A = \begin{bmatrix} 0.9039 & -0.1903 \\ 0.0095 & 0.9990 \end{bmatrix}, B = \begin{bmatrix} 0.0095 \\ 0 \end{bmatrix}$$

que representa un modelo discretizado de un sistema sencillo masa-muelle-amortiguado Figura 2. La solución de DT ARE con los pesos de la función de coste $H_x = \operatorname{diag}(1, 1)$, $H_u = 1$ y $\gamma = 1$ es

$$P_{DARE} = \begin{bmatrix} 5.7529 & 2.5143 \\ 2.5143 & 130.338 \end{bmatrix}$$

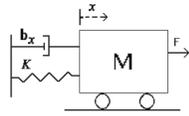


Figura 2: Sistema de masa, muelle y amortiguador

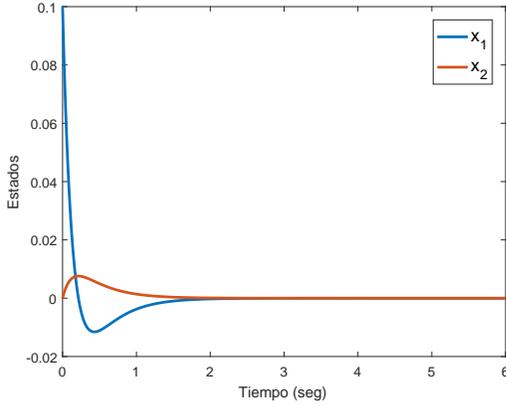


Figura 3: Variables de estado del sistema simulado

Definimos la aproximación de la función de valor considerando un modelo de coste cuadrático en la acción de control:

$$Q(x_k, u_k) = W^T \phi(x_k, u_k)$$

$$\phi(x_k, u_k) = [x_{k1}^2 \quad x_{k1}x_{k2} \quad x_{k1}u_k \quad x_{k2}^2 \quad x_{k2}u_k \quad u_k^2]^T$$

$$W = [w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5 \quad w_6]^T$$

La acción de control óptima es:

$$u_k^* = -\frac{1}{2}w_6^{-1}[w_3 \quad w_5][x_{k1} \quad x_{k2}]^T$$

La implementación online de PI se realizó usando el método de mínimos cuadrados recursivos (*Recursive-Least-Squares, RLS*). Las trayectorias de los estados se muestran en la Figura 3, donde se evidencia como los estados son regulados a cero como es deseable. Los valores finales de la matriz P son:

$$P_{RL} = \begin{bmatrix} 5.7529 & 2.5143 \\ 2.5143 & 130.3380 \end{bmatrix}$$

El algoritmo implementado es un algoritmo adaptativo de control que identifica la función Q a través de técnicas RLS. Para su implementación no se requiere de las matrices de la dinámica del sistema (A, B). El algoritmo efectivamente resuelve la ecuación algebraica de Riccati en línea a tiempo real usando datos $(x_k, u_k, x_{k+1}, u_{k+1})$ medidos en tiempo real en cada instante k . Es necesario agregar ruido de exploración a la señal de control para garantizar una excitación persistente para lograr la convergencia usando RLS.

4.2. Aprendizaje por refuerzo para un sistema lineal discreto usando VI (*Value Iteration*)[7].

En esta simulación se muestra el uso del algoritmo VI para resolver la DT ARE sin conocer la dinámica del sistema. Un modelo lineal puede ser usado para representar el sistema dinámico alrededor de un punto de operación específico con una carga de valor constante. El problema se incrementa con el hecho de que los parámetros de la planta no son conocidos, mientras lo que se busca es una solución de control óptima. El modelo del sistema a considerar aquí es $\dot{x} = Ax + Bu$. Los parámetros del sistema en tiempo continuo se seleccionan aleatoriamente con rango de operación específicos, tenemos

$$A = \begin{bmatrix} -0.0665 & 8 & 0 & 0 \\ 0 & -3.663 & 3.663 & 0 \\ -6.8681 & 0 & -13.7363 & -13.7363 \\ 0.6 & 0 & 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 0 \\ 13.7355 \\ 0 \end{bmatrix}$$

El sistema se ha discretizado con un periodo de muestreo de $T = 0.01s$ y los pesos de la función de costo: $H_x = I$, $H_u = I$, y $\gamma = 1$. La solución es:

$$P_{DARE} = \begin{bmatrix} 0.4805 & 0.4772 & 0.0604 & 0.4771 \\ 0.4772 & 0.7892 & 0.1240 & 0.3855 \\ 0.0604 & 0.1240 & 0.0567 & 0.0304 \\ 0.4771 & 0.3855 & 0.0304 & 2.3509 \end{bmatrix}$$

La implementación de VI se realizó usando mínimos cuadrados por lotes, es así que la ecuación de Riccati es resuelta con los datos $(x_k, x_{k+1}, r(x_k, u_k))$. La aproximación de la función de valor es:

$$Q(x_k, u_k) = W^T \phi(x_k, u_k)$$

$$\phi(x_k, u_k) = [x_{k1}^2, x_{k1}x_{k2}, x_{k1}x_{k3}, x_{k1}x_{k4}, x_{k1}u_k, x_{k2}^2, x_{k2}x_{k3}, x_{k2}x_{k4}, x_{k2}u_k, x_{k3}^2, x_{k3}x_{k4}, x_{k3}u_k, x_{k4}^2, x_{k4}u_k, u_k^2]^T$$

$$W = [w_1 \quad w_2 \quad w_3 \quad \dots \quad w_{14} \quad w_{15}]^T$$

La acción de control óptima es:

$$u_k^* = -\frac{1}{2}w_{15}^{-1}[w_5 \quad w_9 \quad w_{12} \quad w_{14}][x_{k1} \quad x_{k2} \quad x_{k3} \quad x_{k4}]^T$$

Las trayectorias de los estados del sistemas se muestran en la Figura 4, donde se observa que los estados son regulados a cero. Los valores finales de los parámetros estimados para P son

$$P_{RL} = \begin{bmatrix} 0.4753 & 0.4770 & 0.0602 & 0.4769 \\ 0.4770 & 0.7837 & 0.1238 & 0.3852 \\ 0.0602 & 0.1238 & 0.0513 & 0.0302 \\ 0.4769 & 0.3852 & 0.0302 & 2.3457 \end{bmatrix}$$

La Figura 5 muestra la convergencia de P_k a sus valores óptimos P^* durante el proceso de aprendizaje. Para su implementación no se requiere de las matrices de la dinámica del sistema (A, B). El

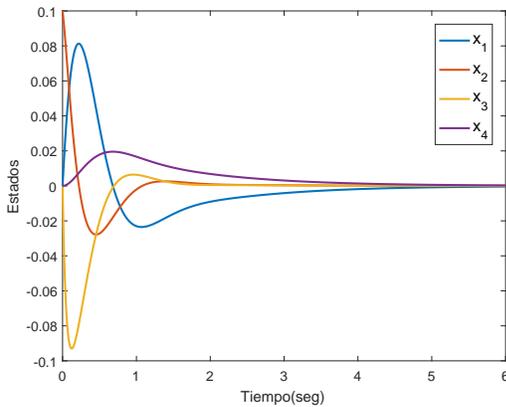


Figura 4: Variables de estado del sistema simulado

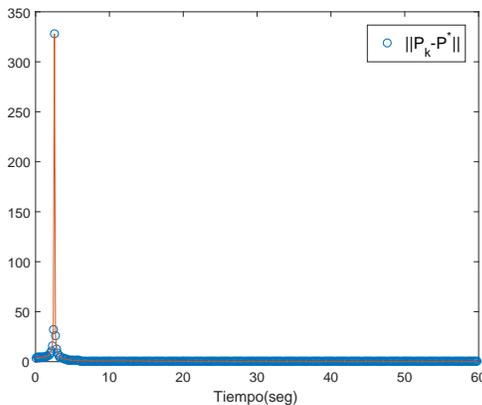


Figura 5: Convergencia de P_k a su valor óptimo P^*

algoritmo efectivamente resuelve la ecuación algebraica de Riccati en línea a tiempo real usando datos $(x_k, u_k, x_{k+1}, u_{k+1})$ medidos en tiempo real en cada instante k . Es necesario agregar ruido de exploración a la señal de control para garantizar una excitación persistente para lograr la convergencia usando LS.

5. Experimentación real: Péndulo de un grado de libertad(1DoF)

En esta sección se describe la implementación práctica del algoritmo de aprendizaje *Q-Learning* en un péndulo de un grado de libertad, con el objetivo de verificar si el algoritmo de aprendizaje nos proporciona una ganancia estabilizante, considerando que no se ha tomado en cuenta para la función de coste las no linealidades existentes en experimentos prácticos. Las implementaciones reales de los algoritmos se han hecho sobre un banco de experimentos mostrado en la Figura 6. El banco de pruebas esta formado por un péndulo de 1DoF cuyo actuador es un motor DC, el péndulo posee un sensor de posición de efecto hall en el primer eslabón. El experimento diseñado se ha definido

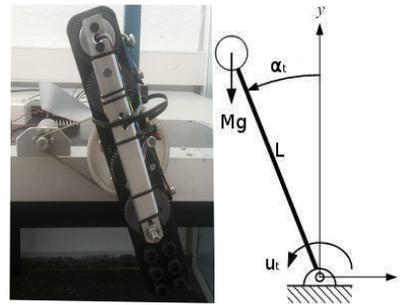


Figura 6: Sistema implementado para pruebas de aprendizaje *Q-Learning*.

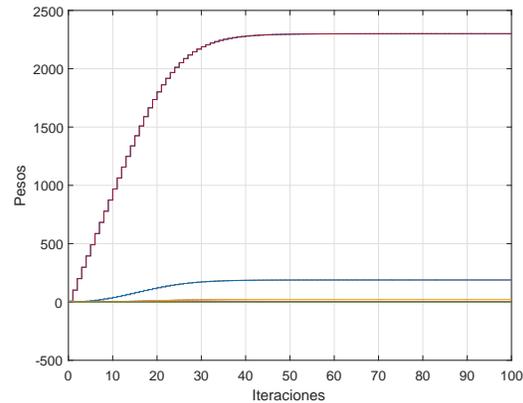


Figura 7: Convergencia de la matriz de parámetros del aprendizaje usando VI.

como la búsqueda del controlador con realimentación del estado para una trayectoria del péndulo desde una posición inicial de $-\frac{\pi}{3}$ [rad] hasta su posición arriba 0 [rad]. En la etapa de exploración se adquirieron 3000 muestras a un periodo de adquisición de $T = 10ms$. Los datos obtenidos en la exploración ingresan al algoritmo de aprendizaje implementado *Q-Learning*, el factor de descuento es $\gamma = 0.98$ y las matrices de ponderación del índice de coste usadas son:

$$H_x = \begin{bmatrix} 100 & 0 \\ 0 & 0.1 \end{bmatrix}, H_u = 0.1$$

La aproximación de la función de valor es $Q(x_k, u_k) = W^T \phi(x_k, u_k)$ y la Figura 7 muestra la convergencia del vector de pesos W . La Figura 8 muestra la evolución de la posición del péndulo usando el controlador aprendido y comparándolo con un controlador proporcional-derivativo(PD).

6. Conclusiones

El aprendizaje por refuerzo proporcionan soluciones a problemas de decisión secuencial, los cuales aparecen en una amplia variedad de campos entre ellos el de los sistemas de control. Gran parte de los algoritmos de aprendizaje por refuerzo se basan en las técnicas de programación dinámica, y la

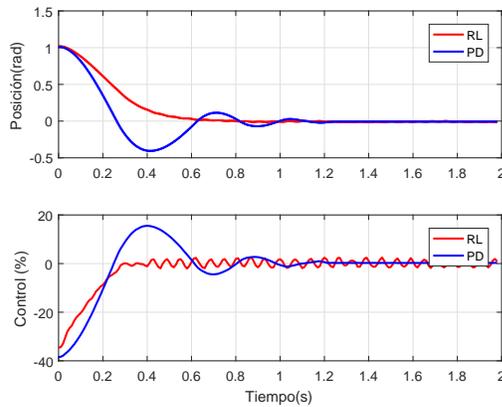


Figura 8: Evolución de la posición del sistema y la acción de control del controlador *Q-Learning* y de un controlador PD implementado desde la condición inicial $x_0 = [-\frac{\pi}{3} \ 0]^T$.

diferencia fundamental yace en que la DP requiere de un modelo de comportamiento mientras que RL no requiere del conocimiento del modelo para proporcionar una solución, específicamente en el caso de los sistemas de control proporcionan un controlador estabilizante y óptimo. El aprendizaje por refuerzo nos permite diseñar controladores adaptativos que convergen a soluciones óptimas usando los datos medidos a lo largo de las trayectorias del sistema. Las simulaciones también permite concluir que la técnica de aprendizaje *Q-Learning* resuelve la ecuación de *Riccati*, en nuestro caso en las simulaciones de manera *online* y en el experimento real de manera *offline*, sin el conocimiento del comportamiento dinámico del sistema, simplemente observando los datos medidos (exploración) a lo largo de las trayectorias del sistema.

7. Agradecimientos

Agradecemos al Ministerio de Economía de España, la Unión Europea DPI2016-81002-R (AEI/FEDER, UE), y al Gobierno de Ecuador (Beca SENESCYT).

Referencias

- [1] Asma Al-Tamimi, Frank L Lewis, and Murad Abu-Khalaf. Model-free q-learning designs for linear discrete-time zero-sum games with application to h-infinity control. *Automatica*, 43(3):473–481, 2007.
- [2] Richard Bellman. *Dynamic programming*. Courier Corporation, 2013.
- [3] Dimitri P Bertsekas and Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA, 1995.

- [4] Dimitri P Bertsekas and Bertsekas. *Neuro-Dynamic Programming*, volume 1. Athena Scientific, 1996.
- [5] Lucian Busoniu, Robert Babuska, Bart De Schutter, and Damien Ernst. *Reinforcement learning and dynamic programming using function approximators*, volume 39. CRC press, 2010.
- [6] Frank L Lewis and Draguna Vrabie. Reinforcement learning and adaptive dynamic programming for feedback control. *Circuits and Systems Magazine, IEEE*, 9(3):32–50, 2009.
- [7] Frank L Lewis, Draguna Vrabie, and Kyriakos G Vamvoudakis. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *Control Systems, IEEE*, 32(6):76–105, 2012.
- [8] Vrabie D. L. Lewis, F. L. and V. L. Syrmos. *Optimal control*. John Wiley & Sons, 2012.
- [9] John J Murray, Chadwick J Cox, George G Lendaris, and Richard Saeks. Adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 32(2):140–153, 2002.
- [10] Warren B Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2011.
- [11] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.
- [12] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.
- [13] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.
- [14] Paul J Werbos. A menu of designs for reinforcement learning over time. *Neural networks for control*, pages 67–95, 1990.
- [15] Paul J Werbos. Approximate dynamic programming for real-time control and neural modeling. *Handbook of intelligent control: Neural, fuzzy, and adaptive approaches*, 15:493–525, 1992.
- [16] Marco Wiering and Martijn van Otterlo. *Reinforcement Learning: State-of-the-art*, volume 12. Springer Science & Business Media, 2012.