Apéndice 12

APRENDIZAJE POR REFUERZO

José Antonio Martín H. y Matilde Santos Peñas

A12.1 INTRODUCCIÓN

En este apéndice se describe un ejemplo de aplicación de la teoría del aprendizaje por refuerzo a un problema clásico de control, el péndulo invertido o Cart-Pole. El péndulo invertido es un sistema dinámico no lineal, ampliamente utilizado como método de prueba para diversos algoritmos de control (reguladores PID, redes neuronales, control borroso, algoritmos genéticos, etc.).

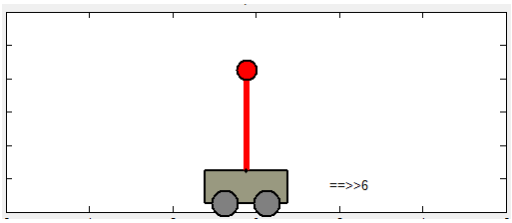

Figura A12.1 El péndulo invertido sobre un coche.

Un péndulo invertido consiste en un péndulo que tiene su masa sobre su punto de pivote. Se configura a menudo con el punto de pivote montado en un coche o plataforma que puede moverse horizontalmente hacia los dos lados (Figura A12.1). Considerando que un péndulo normal es estable al colgar hacia abajo, un péndulo invertido es intrínsecamente inestable y se debe balancear activamente

para lograr mantenerlo vertical, aplicando una fuerza de torsión en el punto de pivote o moviendo el punto de pivote horizontalmente como parte de un sistema de regeneración. La tarea de control consiste entonces en mantener en posición vertical (equilibrada) la varilla con la masa y no dejarla caer, es decir, que el ángulo con la vertical sea cero. Sin embargo, en su aplicación práctica, es necesaria una restricción adicional ya que el coche no puede moverse infinitamente hacia cada lado, de tal forma que se establecen dos topes que determinan la máxima distancia horizontal en la que puede moverse el coche. En caso de topar con uno de los dos extremos se considerará como un episodio fallido.

Así, el sistema a controlar posee 4 variables para describir el estado en el que se encuentra: la posición *x* del coche, la velocidad del coche *x'*, el ángulo θ de la barra con el eje horizontal y la velocidad angular $\dot{\theta}$.

La forma de controlar el péndulo consiste en la aplicación de una fuerza lateral sobre el coche. Esta fuerza que se imprime sobre el coche toma valores discretos en el intervalo [-1.0 a 1.0], con paso de 0.1, lo que da un total de 21 posibles acciones a elegir.

De esta forma, mediante el uso de estas 4 variables que describen el estado del sistema y las 21 acciones disponibles, se puede definir el problema de control del péndulo invertido como un proceso de decisión de Markov (MDP) y utilizar un algoritmo de Aprendizaje por Refuerzo, por ejemplo el algoritmo SARSA (Algoritmo 11.3) descrito en el capítulo 11, para controlarlo y mantenerlo en una posición de estabilidad.

A12.2 SIMULACIÓN EN SOFTWARE

Para demostrar el uso del Aprendizaje por Refuerzo en esta tarea de control se presenta un ejemplo desarrollado con el software de computación Matlab. La aplicación consta de una ventana que implementa la interfaz de usuario, figura A12.2:

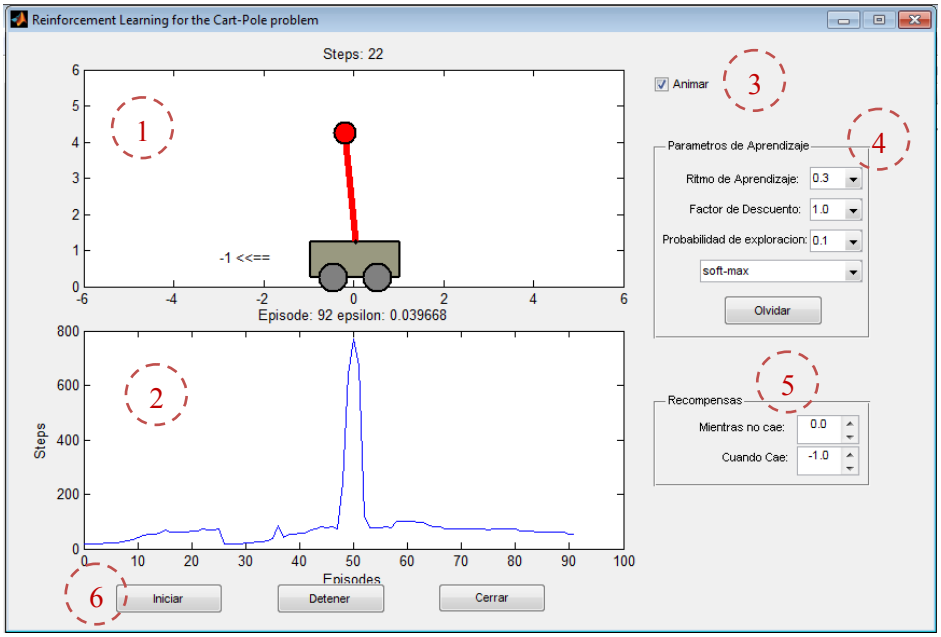

Figura A12.2. Interfaz de usuario de la aplicación de aprendizaje por refuerzo en el péndulo invertido desarrollada en Matlab.

La interfaz de usuario se despliega cuando se inicia la aplicación desde la consola de comandos de Matlab, mediante la invocación al procedimiento "*PenduloInvertido*", es decir:

>>PenduloInvertido <Pulsar Intro>

En la Figura A12.2 se han etiquetado con un número y un círculo rojo las diferentes partes de la interfaz de usuario:

1. Área de dibujo de la simulación.
2. Gráfica de evolución del aprendizaje.
3. Botón animar (mostrar animación).
4. Parámetros de Aprendizaje.
5. Especificación de la Recompensa.
6. Área de control.

1. El área de dibujo (ventana superior) muestra, al pulsar el botón Animar (3), el gráfico del péndulo invertido sobre un vehículo. Esta imagen está animada, correspondiendo su movimiento al funcionamiento del sistema en ejecución. Permite observar de forma gráfica y directa la forma en la cual va aprendiendo el agente a controlar el péndulo.

En esta ventana se muestran los límites del desplazamiento lateral del vehículo, fijados a [-4,4], así como la dirección del movimiento del vehículo y un número que indica la fuerza aplicada que va desde -10 a 10 Newton.

Sobre este recuadro aparece actualizado de forma continua el número de pasos correspondientes al episodio actual. Un episodio exitoso se simula durante 1000 pasos. Es decir, el péndulo se habrá mantenido en equilibrio durante 1000 pasos, iteraciones simples del sistema dinámico que comprende la simulación del péndulo invertido.

Cuando el péndulo pierde la posición de equilibrio o el vehículo alcanza alguno de los topes laterales, el episodio se considera fallido y aunque el número de pasos no haya llegado a 1000, se genera un nuevo episodio, ver Figura A12.3.

El botón animar, que permite ver el dibujo en movimiento del péndulo, ralentiza la ejecución de los episodios. También es posible simular el sistema y aplicar el algoritmo de aprendizaje sin mostrar el gráfico del péndulo invertido, pudiéndose observar la información numérica correspondiente a la simulación (número de pasos) así como la gráfica del aprendizaje, pasos frente a episodios, en la ventana inferior.

2. La evolución del aprendizaje se muestra en la ventana inferior de la interfaz. En ella aparece la gráfica de la relación entre el tiempo que permanece el péndulo en equilibrio y el número de episodios que se han experimentado hasta el momento. El número de episodio se muestra tanto numéricamente sobre la ventana como en el eje inferior de la gráfica.

Junto con el número de episodios aparece el valor del parámetro “épsilon” que irá decreciendo en cada episodio para asegurar la convergencia. Inicialmente su valor es de 0.1 y puede también ser fijado por el usuario, parámetro probabilidad de exploración. En el ejemplo de la Figura A12.3 su valor inicial es 0.1.

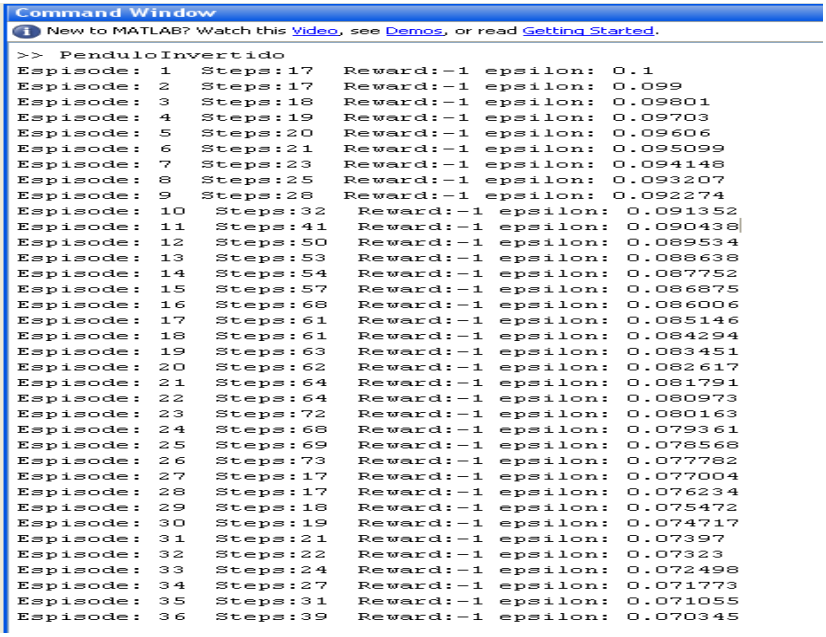

Figura A12.3 Ventana de comandos donde aparece la evolución del algoritmo de aprendizaje por refuerzo.

3. En la región de parámetros del aprendizaje se puede distinguir 5 elementos:

- Ritmo de aprendizaje (α) cuyos valores pueden seleccionarse de una lista desplegable cuyo rango está comprendido entre 0.0 y 1.0 con paso 0.1, es decir, 11 valores posibles donde 0.0 significa que el sistema no aprende. Como se expuso en la teoría, este parámetro puede interpretarse como el ritmo de aprendizaje de forma análoga a como se utiliza en el algoritmo de retro-propagación para redes de perceptrones multicapa. El máximo valor es 1.0.
- Factor de descuento (γ), cuyos valores pueden seleccionarse de una lista desplegable cuyo rango está comprendido entre 0.0 y 1.0 con paso 0.1. Este parámetro, en la medida en que sea menor que 1, hará que el aprendizaje tome menos en consideración la recompensa futura.
- Probabilidad de exploración (ε) cuyos valores pueden seleccionarse de una lista desplegable donde ε está comprendido entre 0.0 y 1.0 con paso 0.1. Como se ha indicado, su valor va decreciendo para asegurar la convergencia.
- Método de Exploración: puede seleccionarse uno de los mecanismos implementados mediante una lista desplegable. Los posibles mecanismos de

selección de acciones disponibles en la aplicación son el ϵ-greedy y el soft-max.

- Botón olvidar: al pulsar este botón se reinicializa la memoria, tabla con todos los valores *Q* del algoritmo, eliminando toda la información aprendida de la experiencia de tal forma que el algoritmo comenzaría a aprender de nuevo desde el inicio.

4. En la región de la recompensa hay dos cajas de texto donde se especifican los valores que recibirá el agente para dos posibles casos:

a) Mientras el péndulo esté en equilibrio o no se llegue a un tope lateral. Puede establecerse cualquier valor real. Por defecto tiene el valor 0 indicando que no hay penalización por permanecer en este estado.

b) Cuando se pierde el equilibrio y el péndulo cae, o bien el vehículo alcanza un tope lateral, entonces ese episodio ha sido fallido. Por defecto la recompensa (castigo en este caso) es -1 para indicar al sistema que este estado no es deseable, figura A12.3. Puede establecerse cualquier valor real.

Como puede verse, las recompensas guían el aprendizaje en el sentido de que son éstas las que expresan el objetivo o meta en el problema.

5. En la región inferior de la interfaz se muestran tres botones de control, desde donde se controla el flujo del programa de simulación. Estos botones permiten iniciar la simulación, cerrarla o detenerla. En este último caso, al volver a pulsar iniciar la ejecución se reanuda donde se había quedado.

Por último, se muestra una ejecución, figura A12.4, donde se puede observar que el sistema ha aprendido. Los episodios son ahora de 1000 pasos, lo que se muestra tanto en la ventana de comandos de Matlab como en la gráfica del aprendizaje. En esta última se puede observar cómo inicialmente el número de pasos por episodio era muy bajo. Se produce un pico de episodios de hasta casi 800 pasos, pero luego vuelve a decaer. Por último, empiezan a sucederse episodios de 1000 pasos, es decir, exitosos, hasta que a partir del episodio 130 el sistema ha aprendido y ya la gráfica del número de pasos por episodio se mantiene constante al máximo valor. La recompensa en esos casos positivos es por defecto cero, como puede comprobarse en la línea de comandos.

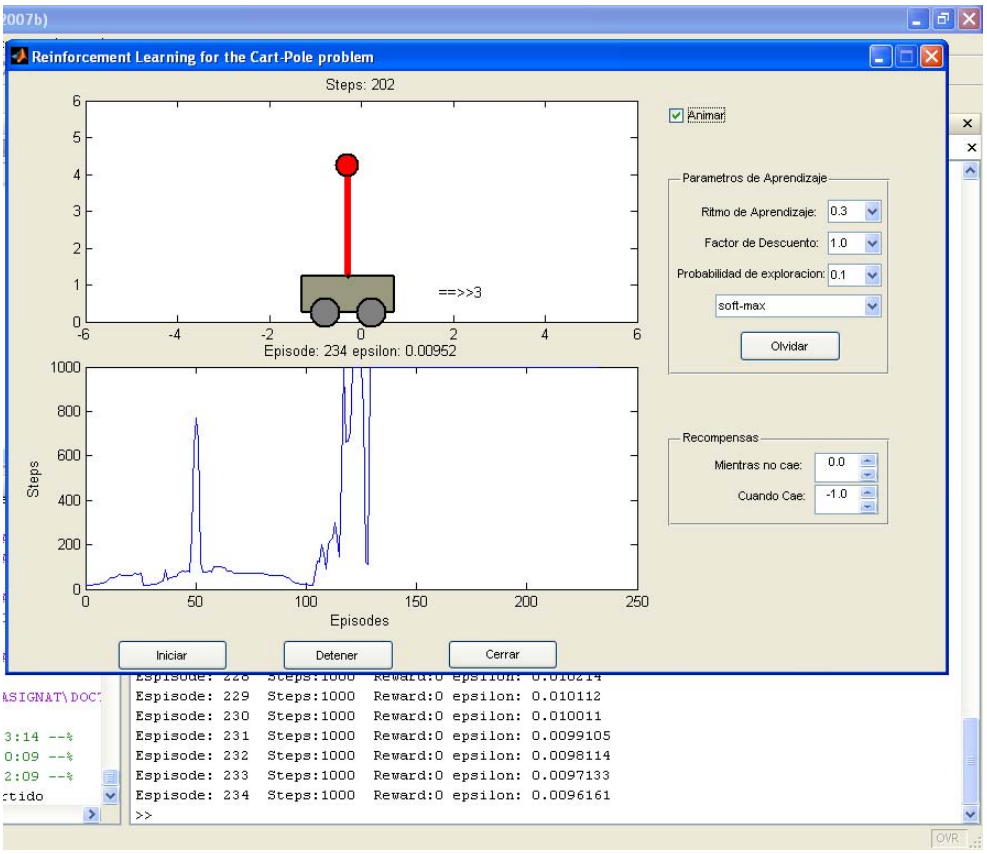

Figura A12.4 Resultados de una simulación en la que el sistema ha aprendido.