Revista Iberoamericana de Ingeniería Mecánica. Vol. 9, N.º 2, pp. 77-91, 2005

ALGORITMO DE APRENDIZAJE POR REFUERZO CONTINUO PARA EL CONTROL DE UN SISTEMA DE SUSPENSIÓN SEMI-ACTIVA

Mª JESÚS LÓPEZ BOADA, BEATRIZ LÓPEZ BOADA, VICENTE DÍAZ LÓPEZ

Universidad Carlos III de Madrid
Departamento de Ingeniería Mecánica
Avda. de la Universidad 30, 28911 Leganés, Madrid, España

(Recibido 15 de junio de 2004, para publicación 8 de noviembre de 2004)

Resumen – Este artículo presenta un algoritmo de aprendizaje por refuerzo utilizando redes neuronales que permite a un vehículo equipado con suspensión semi-activa mejorar continuamente tanto el confort de los pasajeros como el contacto neumático-suelo cuando atraviesa terrenos regulares y/o irregulares. La arquitectura de red neuronal implementada permite trabajar con espacios de entradas y salidas continuas, tiene una buena resistencia a olvidar lo aprendido anteriormente y aprende rápidamente. Otras ventajas que dicho algoritmo presenta son, por una parte, que no es necesario estimar una recompensa esperada porque el controlador recibe una señal de refuerzo real continua cada vez que realiza una acción y, por otra parte, que el sistema aprende *on-line*, por lo que el sistema puede adaptarse a cambios producidos en el entorno y en el propio sistema de suspensión. El controlador neuronal propuesto ha sido estudiado utilizando un modelo de un cuarto de vehículo. Los resultados de simulación obtenidos muestran la efectividad del algoritmo propuesto.

1. INTRODUCCIÓN

En los últimos años, la mejora de la seguridad en la conducción y el confort de los viajeros son una parte central dentro de las investigaciones que se está realizando dentro del campo de los vehículos automóviles. El sistema de suspensión de un vehículo influye de manera importante en estos dos parámetros. El sistema de suspensión debe soportar el peso del vehículo y proporcionar tanto un control de la dirección adecuado durante las maniobras realizadas para garantizar el contacto neumático-suelo, como aislar a los pasajeros de las vibraciones debidas a las perturbaciones de la carretera. Un buen confort durante la conducción requiere una suspensión blanda, sin embargo, este tipo de sistemas producen un balanceo excesivo en las curvas y, además, favorece el cabeceo durante la frenada. Esto puede resultar peligroso para la seguridad de los pasajeros. La solución a esta falta de seguridad sería la utilización de una suspensión más rígida, que controlaría mejor el balanceo en curva y el cabeceo pero que sin embargo, disminuiría el confort de los pasajeros. Este problema se resuelve por un sistema de suspensión que permita adaptarse a las características del terreno.

Los sistemas de suspensión se clasifican en pasivos, semi-activos y activos [1][2]. Los sistemas de suspensión pasivos están formados por elementos pasivos como muelles y amortiguadores que no introducen energía al sistema. Los sistemas semi-activos actúan modificando la rigidez del muelle o del amortiguador mediante actuadores. En los sistemas de suspensión activa, los elementos mecánicos (muelle y amortiguador) son reemplazados por un elemento activo (actuador hidráulico) que es capaz de generar fuerzas de acuerdo a un algoritmo de control.

Para el estudio del comportamiento de los sistemas de suspensión semi-activa y activa se han utilizado diferentes modelos de vehículos. En [3], [4] y [5] se utiliza un modelo de vehículo completo (7 grados de libertad). En [6] se emplea un modelo de vehículo de 4 grados de libertad. En [7] se utiliza un sistema no-lineal de 6 grados de libertad. Mazaruddin et al. [8] presentan un método para identificar modelos de sistemas de suspensión utilizando técnicas neuro-fuzzy. Sin embargo, la mayoría de los autores [9][10][21][23] utilizan un modelo de un cuarto de vehículo debido a que este modelo tiene en cuenta muchas de las características importantes de la suspensión manteniendo la simplicidad del modelo.

En la literatura, se han propuesto diferentes controladores para el control de diferentes tipos de sistemas de suspensión: semi-activa [11][12][13][14] y activa [10][15][16]. Muchos autores, utilizan algoritmos no-lineales para el control de la suspensión de un vehículo tales como control borroso o *fuzzy* [17][18][19][20], control en modo deslizante [4][6], algoritmos genéticos [21], controlador neuronal [22], control adaptativo ARMAKOV [23], etc. Las principales ventajas de utilizar un control no-lineal es que se basa en modelos mas reales consiguiendo que el sistema mejore su comportamiento. Muchos de estos controladores no-lineales permiten al sistema de suspensión adaptarse o aprender a mejorar su comportamiento durante su funcionamiento.

En función de la información recibida durante el aprendizaje, los métodos de aprendizaje se pueden clasificar en aprendizaje supervisado y no supervisado [24]. Los métodos de aprendizaje supervisado se caracterizan porque existe un *maestro* que proporciona las salidas deseadas para cada vector de entradas. Estos métodos son muy potentes debido a que trabajan con mucha información. Los inconvenientes que presentan son que el aprendizaje se realiza *off-line* y es necesario conocer con detalle cómo se tiene que comportar el sistema. En los métodos de aprendizaje no supervisado no existe un *maestro* que instruya sobre cuales han de ser las salidas correctas ante determinadas entradas. Dentro de estos métodos se encuentra el aprendizaje por refuerzo [25]. En este caso, existe un *crítico* que proporciona una información más evaluativa que instruccional. Las ventajas que ofrece son, en primer lugar, que no necesita un conocimiento completo acerca del sistema y, en segundo lugar, que puede ser utilizado *on-line*, el sistema puede mejorar continuamente su comportamiento mientras que aprende. Howell et al. [26] utilizan el aprendizaje por refuerzo para mejorar el confort durante la conducción de un vehículo experimental equipado con suspensión semi-activa minimizando el error cuadrático medio de la aceleración vertical del vehículo.

En este artículo, se presenta un algoritmo de aprendizaje por refuerzo utilizando redes neuronales que permite a un vehículo equipado con suspensión semi-activa mejorar continuamente no solamente el confort en la conducción sino también el contacto neumático-suelo tanto en terrenos regulares como irregulares.

2. MODELO DE LA SUSPENSIÓN DE UN VEHÍCULO

Para el estudio del comportamiento de la dinámica de un sistema de suspensión semi-activa se ha utilizado el modelo de un cuarto de vehículo como se muestra en la Figura 1 [27][28] debido a que este modelo tiene en cuenta muchas de las características importantes de la suspensión manteniendo la simplicidad del modelo. Un actuador se conecta en paralelo con el muelle y amortiguador de la suspensión pasiva con objeto de controlar la suspensión del vehículo y mejorar su comportamiento.

El neumático es modelado por un simple muelle con rigidez k_2. La masa del neumático y masa no suspendida se define como m_u. Se considera que el neumático está en contacto permanente con el suelo. La masa suspendida del vehículo (m_s) se considera como un cuerpo rígido. El muelle (k_1), el amortiguador (C_s) y el actuador situados entre la masa no suspendida (m_u) y masa suspendida (m_s) constituyen el sistema de suspensión semi-activa.

Las ecuaciones de la dinámica del modelo de suspensión de un cuarto de vehículo son:

- masa suspendida

$$m_s \cdot \ddot{z}_s = k_1 \cdot (z_w - z_s) + C_s(\dot{z}_w - \dot{z}_s) - f_a \tag{1}$$

- masa no suspendida

$$m_u \cdot \ddot{z}_w = -k_1 \cdot (z_w - z_s) - C_s \cdot (\dot{z}_w - \dot{z}_s) - k_2 \cdot (z_w - z_r) + f_a \tag{2}$$

A partir de las ecuaciones (1) y (2) se puede formular las siguientes ecuaciones en el espacio de estados:

$$\dot{X} = A \cdot X + B \cdot u \tag{3}$$

$$Y = C \cdot X + D \cdot u \tag{4}$$

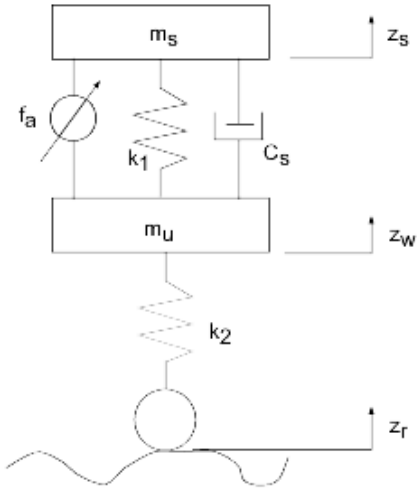

Fig. 1. Modelo de un cuarto de vehículo.

donde X es el vector de variables independiente, u es el vector de entradas e Y es el vector de salidas:

$$X=\begin{bmatrix} z_s \\ z_w \\ \dot{z}_s \\ \dot{z}_w \end{bmatrix};\quad A=\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -\frac{k_1}{m_s} & \frac{k_1}{m_s} & -\frac{C_s}{m_s} & \frac{C_s}{m_s} \\ \frac{k_1}{m_u} & -\frac{k_1+k_2}{m_u} & \frac{C_s}{m_u} & -\frac{C_s}{m_u} \end{bmatrix};\quad B=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ -\frac{1}{m_s} & 0 \\ \frac{1}{m_u} & \frac{k_2}{m_u} \end{bmatrix} \tag{5}$$

$$u=\begin{bmatrix} f_a \\ z_r \end{bmatrix}\quad y=x\quad C=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}\quad D=\begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{6}$$

El vector de entradas está formado por la fuerza del actuador (f_a) y las perturbaciones de la carretera (z_r).

Para el diseño y evaluación de un sistema de suspensión de un vehículo se tienen en cuenta tres parámetros [28]: 1) el aislamiento a las vibraciones de la masa suspendida que determina el confort de los pasajeros, 2) la carrera de la suspensión que limita el movimiento del cuerpo del vehículo y 3) el contacto neumático-suelo que proporciona las fuerzas laterales y de frenado adecuadas. Para obtener dichos parámetros las siguientes variables son examinadas: 1) la aceleración vertical de la masa suspendida ($\ddot{z}_s$), 2) la deflexión de la suspensión ($z_w - z_s$), y 3) la deflexión del neumático ($z_r - z_w$).

3. APRENDIZAJE POR REFUERZO

El aprendizaje por refuerzo, RL (*Reinforcement Learning*), consiste en mapear situaciones a acciones maximizando un escalar denominado señal de refuerzo o recompensa [25][29]. Es una técnica de aprendizaje basada en prueba y error. El aprendizaje por refuerzo se utiliza cuando no existe una información detallada sobre la salida deseada. A diferencia del aprendizaje supervisado, no existe un maestro que instruya sobre las salidas correctas ante determinadas entradas, pero sí un *crítico* que proporciona una información más evaluativa que instruccional.

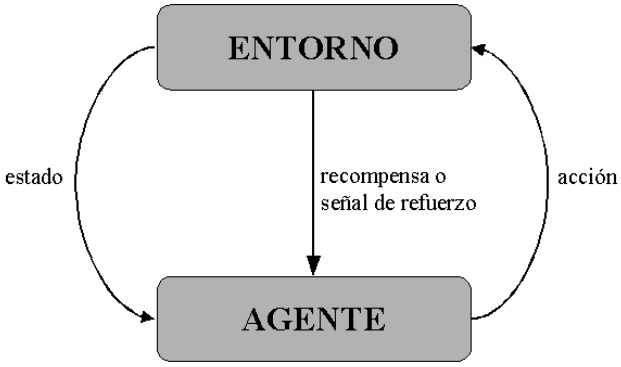

Fig. 2. Interacción entre los elementos de un sistema de aprendizaje por refuerzo.

En este tipo de aprendizaje no se indican cuales son las acciones correctas. La idea es que el sistema explore el entorno y observe el resultado de las acciones sobre algún índice de resultados que permita obtener información para su aprendizaje.

Los elementos que forman un sistema de aprendizaje por refuerzo son un *agente*, el *entorno*, una *política,* una *función de valor*, y, opcionalmente, un *modelo* del entorno (ver Figura 2). El agente es el que aprende y toma las decisiones. El *entorno* es lo que interacciona con el agente. El *agente* selecciona acciones y el entorno responde a estas acciones presentando nuevas situaciones al agente. Una *política* define la forma de comportarse del agente. Una política es una correspondencia de estados percibidos del entorno a acciones a ser llevadas a cabo. Esto corresponde a lo que en psicología se llama un conjunto de reglas o asociaciones estímulo-respuesta. En algunos casos, esta política puede ser una simple función, mientras que en otros casos puede implicar cálculos más complejos tales como procesos de búsqueda. La política es el corazón del agente en el sentido de ser suficiente para determinar su comportamiento. En general, las políticas son estocásticas. Una *función de recompensa* define el objetivo o la meta en el aprendizaje por refuerzo. Esta función mapea cada estado percibido del entorno a un valor, llamado recompensa, indicando la conveniencia del estado. El objetivo único del agente en el aprendizaje por refuerzo es maximizar la recompensa total recibida. La función de recompensa define lo que es bueno y malo para el agente y no puede ser modificable por éste. En general, las funciones de recompensa son estocásticas. Mientras que la función de recompensa indica lo que es bueno en un sentido inmediato, una *función de valor*, indica lo que es bueno a largo plazo. El valor de un estado es la cantidad total de recompensa que un agente puede esperar acumular sobre el futuro, empezando en ese estado. Por último, el *modelo* trata de imitar el comportamiento del entorno. Por ejemplo, dado un estado y una acción, el modelo debería predecir el siguiente estado resultante y la recompensa obtenida.

Un agente de aprendizaje tiene que explorar el entorno con objeto de adquirir conocimiento y seleccionar mejores acciones en el futuro. Para obtener una buena recompensa, el agente tiene que seleccionar, entre las acciones que ha ejecutado anteriormente, aquella que obtenga una recompensa más alta. Pero, al mismo tiempo, para descubrir estas acciones tiene que ejecutar acciones no seleccionadas anteriormente. El agente tiene *explotar* aquellas acciones que ya conoce para obtener una estimación fiable de sus recompensas esperadas y tiene que *explorar* también con el fin de seleccionar mejores acciones en el futuro. Por tanto, el agente tiene que intentar una variedad de acciones y, progresivamente, favorecer aquellas que parezcan mejores. Este problema se denomina equilibrio entre exploración y explotación. Diferentes autores han propuestos diferentes estrategias para resolver dicho problema [30].

El aprendizaje por refuerzo implica dos problemas: problema de asignación de crédito temporal y problema de asignación de crédito estructural [31]. El problema de asignación de crédito temporal aparece debido a que la señal de refuerzo o recompensa puede ser recibida retrasada en el tiempo. La señal de refuerzo informa sobre el éxito o fallo alcanzado después de que una secuencia de acciones han sido realizadas. El problema es cómo asignar a cada acción la recompensa obtenida. Para resolver este problema algunos algoritmos de aprendizaje por refuerzo se basan en estimar una recompensa esperada o predecir evaluaciones futuras como son las diferencias temporales TD(λ) [32]. Dentro de estos algoritmos se en-

cuentran el algoritmo Crítico Heurístico Adaptativo *AHC* (*Adaptive Heuristic Critic*) [33] y el algoritmo Q'Learning [34]. El problema de asignación de crédito estructural aparece cuando el sistema que está aprendiendo está formado por más de un componente y las acciones realizadas dependen de varios de estos componentes. En estos casos, la recompensa recibida por el sistema para una acción determinada debe ser correctamente distribuida entre los diferentes componentes que intervinieron en la misma. Para resolver este problema diferentes métodos han sido propuestos como el métodos del gradiente y métodos basados en el principio de cambio mínimo [35][36].

El aprendizaje por refuerzo ha sido aplicado a diferentes áreas como redes de computadores [37], robótica [38][39][40], control de potencia [41], vehículos de carretera [42], control de tráfico [43], etc. mostrando buenos resultados.

4. ALGORITMO DE APRENDIZAJE POR REFUERZO CONTINUO

En muchos de los algoritmos mencionados en la sección anterior, la señal de refuerzo informa únicamente de si el sistema ha alcanzado el objetivo o no. En estos casos, la señal de refuerzo es un escalar binario cuyos valores típicos son 0 ó 1 (0 si el sistema sigue funcionando y 1 si el sistema ha alcanzado el objetivo), y puede estar retardada en el tiempo. El éxito del proceso de aprendizaje depende principalmente de cómo se defina la señal de refuerzo externa y de cuándo el sistema la reciba. Cuanto mayor sea el tiempo en el que el sistema recibe el refuerzo, mayor será el tiempo que tarde el sistema en aprender. En este trabajo, se propone un algoritmo de aprendizaje por refuerzo que recibe una señal de refuerzo cada vez que el sistema realiza una acción. Este refuerzo es una señal continua comprendida entre (0, 1) y permite valorar cómo de bien el sistema ha realizado la acción. Con esto se consigue que la velocidad de aprendizaje aumente. En este caso, el sistema puede comparar el resultado de la acción realizada con el resultado de la última acción realizada en el mismo estado. En este caso, no es necesario estimar una recompensa esperada.

Muchos de los algoritmos de aprendizaje por refuerzo trabajan con entradas y salidas discretas. Sin embargo, muchas aplicaciones requieren trabajar con espacios continuos definidos por medio de variables continuas. Uno de los problemas que se presentan al trabajar con espacios de entradas continuos es cómo afrontar el número infinito de estados percibidos. Un método generalizado es discretizar el espacio de entradas en regiones en la que dentro de cada una de estas regiones las entradas se asocian a la misma acción [44][45][46]. Por otra parte, los inconvenientes de trabajar con espacios de acciones discretos son, que posibles acciones buenas no se den y que, además, el control es menos fino. Cuando el espacio de acciones es discreto, la implementación del aprendizaje por refuerzo es sencilla, el sistema tiene que elegir una acción de un conjunto fijo y finito de acciones, siendo esta acción la que proporcione una mayor recompensa. Cuando la acción es continua el problema no es tan obvio debido a que el conjunto de posibles acciones es infinito. Para resolver este problema algunos autores añaden ruido a las acciones propuestas [31][47][48].

En algunos algoritmos de aprendizaje por refuerzo, se utilizan redes neuronales para su implementación. Las ventajas que ofrecen son su flexibilidad, robustez al ruido y capacidad de adaptación. En este trabajo se propone un algoritmo de aprendizaje por refuerzo que permite a un sistema de suspensión semiactiva de un vehículo mejorar su comportamiento *on-line* tanto en terrenos regulares como irregulares. La red neuronal implementada trabaja con espacios de entradas y salidas continuos y con señal de refuerzo continua. Anteriormente, dicho algoritmo ha sido utilizado en un robot móvil autónomo para aprender diferentes habilidades como *Seguimiento Visual* [49]*, Ir a un Punto, Seguimiento de Contorno a la Izquierda y Derecha* [50], mostrando muy buenos resultados.

4.1. Arquitectura de la red neuronal

La arquitectura de la red neuronal propuesta para implementar el algoritmo de aprendizaje por refuerzo propuesto consta de dos capas como se muestra en la Figura 3.

La capa de entrada está formada por nodos RBF (*Radial Basis Function*) y se encarga de discretizar el espacio de entrada a la red. Los nodos RBF permiten una transición más suave entre un estado y otro. La función de activación de cada nodo RBF tiene una influencia local permitiendo que el conocimiento ad-

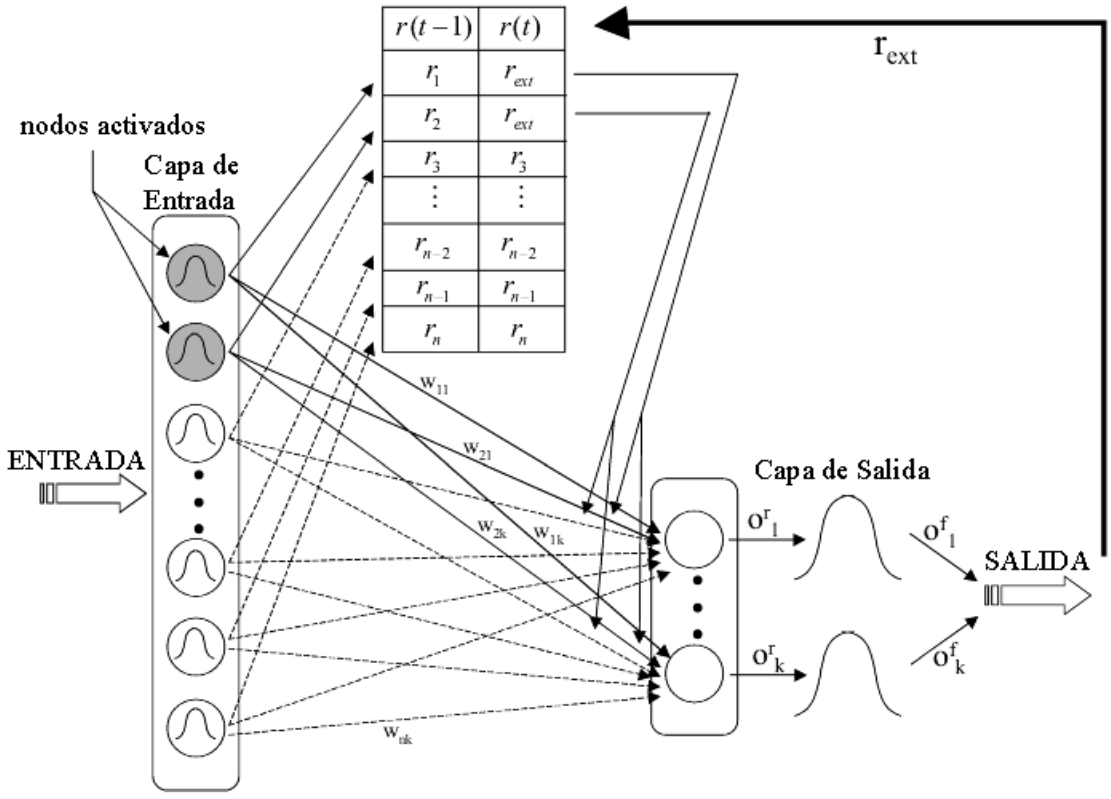

Fig. 3. Estructura de la red neuronal utilizada para implementar el algoritmo de aprendizaje por refuerzo.

quirido en una región determinada influya poco o nada en lo aprendido en otras regiones. El valor de activación de cada nodo depende de cómo de próximo se encuentre el vector de entradas a su centro. Si la activación es 0 indica que la situación percibida está fuera del campo receptivo del nodo. Si el valor de activación es 1 significa que la situación percibida coincide con el centro del nodo. Las entradas presentadas a la red son previamente normalizadas permitiendo que todas participen en el aprendizaje con el mismo peso.

La capa de salida está formada por unidades estocásticas lineales. Con esto se consigue una búsqueda de mejores acciones en el espacio de salida. Cada nodo de la capa de salida representa un acción.

4.1.1. Capa de entrada

El espacio de entrada se divide en regiones discretas utilizando nodos RBF. El valor de activación de cada nodo j es:

$$i_j = \exp^{-\frac{\|\vec{i}-\vec{c}_j\|^2}{\sigma_{rbf}^2}}$$

donde $\vec{i}$ es el vector de entradas presentado a la red, $\vec{c}_i$ es el centro de cada nodo y σ_{rbf} es el ancho de la función de activación. Una vez obtenidos los valores de activación de cada nodo, éstos se normalizan:

$$i_j^n = \frac{i_j}{\sum_{k=1}^{n_n} i_k}$$

donde n_n es el número de nodos creados. Los nodos se crean dinámicamente a medida que se necesitan. Existen nodos únicamente en aquellas situaciones donde el sistema ha explorado. Esto permite que el tamaño de la red sea lo más pequeño posible. Cada vez que una situación se presenta a la red, se calcula el valor de activación para cada nodo. Si todos los valores de activación obtenidos están por debajo de un umbral, $a_{\min}$, se crea un nuevo nodo cuyo centro coincide con el vector de entradas presentado a la red, $\vec{c}_i = \vec{i}$. Los pesos asociados a las conexiones entre el nuevo nodo creado y la capa de salida son inicializados a valores pequeños y aleatorios.

4.1.2. Capa de salida

La capa de salida debe encontrar la mejor acción para cada situación presentada a la red. La acción que la capa de salida de la red neuronal recomienda es la suma ponderada de los valores obtenidos en la capa de entrada:

$$o_k^r = \sum_{j=1}^{n_n} w_{jk} \cdot i_j^n \qquad 1 \le k \le n_0$$

donde n_0 es el número de nodos de la capa de salida. Durante el proceso de aprendizaje, la red necesita explorar todas las posibles acciones para una misma situación presentada con el objetivo de descubrir cual es la mejor. Esto se consigue añadiendo ruido a la salida recomendada. En la red neuronal propuesta, la acción final se obtiene a partir de una distribución normal centrada en la acción recomendada y con varianza σ:

$$o_k^f = N(o_k^r, \sigma)$$

A medida que el sistema aprende, el valor de σ se reduce, de esta manera, el sistema puede realizar la misma situación aprendida [50].

Durante el proceso de aprendizaje de la red los pesos de la capa de salida se adaptan de acuerdo con las ecuaciones siguientes:

$$w_{jk}(t+1) = w_{jk}(t) + \Delta w_{jk}(t)$$

$$\Delta w_{jk}(t) = \beta \cdot (r_{j'}(t) - r_{j'}(t-1)) \cdot \frac{\mu_{jk}(t)}{\sum_l \mu_{lk}(t)} \;\Big|\; j' = \arg\max_j i_j^n$$

$$e_{jk}(t) = \frac{o_k^f - o_k^r}{\sigma} i_j^n$$

$$\mu_{jk}(t+1) = \upsilon \cdot \mu_{jk}(t) + (1-\upsilon) \cdot e_{jk}(t)$$

donde β es la velocidad de aprendizaje, μ_{jk} y e_{jk} son la traza de elegibilidad y la elegibilidad del peso w_{jk} respectivamente, y υ es un valor comprendido entre [0, 1]. La elegibilidad del peso w_{jk}, permite castigar o recompensar no sólo a la última acción sino también a las anteriores. Los valores de $r_{j'}$ corresponden a los refuerzos asociados al nodo de la capa de entrada con máxima activación en ese instante de tiempo. Los valores de refuerzo r_j asociados a cada peso se obtienen a partir de la expresión:

$$r_j(t) = \begin{cases} r_{ext}(t) & \text{Si } i_j^n \neq 0 \\ r_j(t-1) & \text{de otro modo} \end{cases}$$

donde r_{ext} es el refuerzo externo obtenido por la red neuronal. Los resultados de las acciones dependen de los estados activos cuando dichas acciones son ejecutadas. Por lo tanto, únicamente se actualizarán los valores de los refuerzos asociados a esos estados.

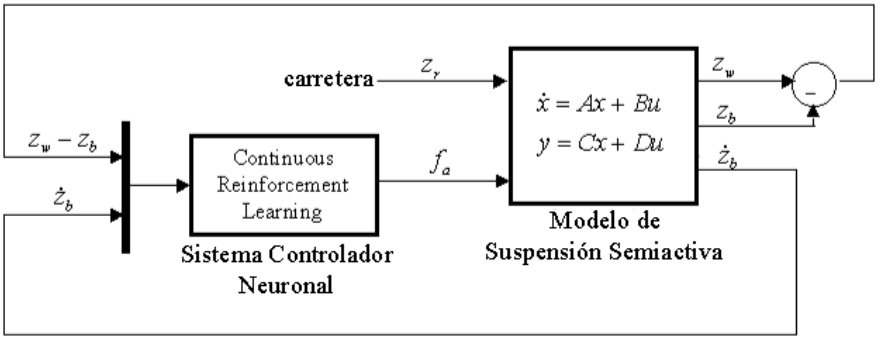

Fig. 4. Estructura del sistema controlador neuronal.

5. RESULTADOS DE SIMULACIÓN

Los resultados de simulación han sido obtenidos utilizando un modelo de un cuarto de vehículo equipado con suspensión semi-activa. El software empleado para la simulación es la herramienta SIMULINK de MATLAB©. La estructura del sistema controlador neuronal propuesto se muestra en la Figura 4. Las tres variables obtenidas a partir del modelo de suspensión semi-activa que son utilizadas por el controlador neuronal para determinar la fuerza del actuador son: la deflexión de la suspensión ($(z_w - z_b)(t)$), el error de la deflexión de la suspensión ($(z_w - z_b)(t)$- $(z_w - z_b)(t\text{-}1)$) y la velocidad vertical de la masa suspendida ($\dot{z}_b(t)$).

El algoritmo de aprendizaje por refuerzo está escrito en código C y ha sido incorporado al modelo de SIMULINK utilizando una *S-function*. La Tabla 1 muestran los valores de los parámetros de la suspensión utilizados y que corresponden a valores típicos de la suspensión de un vehículo.

Para comprobar la fiabilidad del algoritmo de aprendizaje por refuerzo propuesto para el control de un sistema de suspensión semi-activa de un vehículo han sido utilizados dos perfiles de carretera. El primer perfil de carretera corresponde a un terreno regular y es representado mediante un pulso cuadrado de amplitud 0.1 metros, periodo de 6 segundos y ancho del pulso un 50% del periodo (ver Figura 5.a). El segundo perfil corresponde a un terreno irregular y es representado mediante una secuencia de 4 segundos como se muestra en la Figura 5.b.

La señal de refuerzo externa que el controlador neuronal recibe es:

$$r_{ext} = e^{-2.0(z_w - z_b)^2} \cdot e^{-0.5(error_{z_w - z_b})^2} \cdot e^{-2.0(\dot{z}_b)^2}$$

En este caso, el refuerzo es máximo cuando la deflexión de la suspensión, el error de la deflexión de la suspensión y la velocidad vertical de la masa suspendida se hacen cero. Las Figuras 6, 7 y 8 comparan la respuesta de un sistema de suspensión pasiva (línea negra), con la respuesta del sistema de suspensión semi-activa (línea gris) en un terreno de perfil regular después de 1153 ciclos de aprendizaje (1 ciclo de aprendizaje corresponde a un periodo de 6 segundos). Los resultados obtenidos muestran que la deflexión de la suspensión y la aceleración vertical de la masa suspendida en el sistema de suspensión semi-activa

Tabla 1. Valores típicos de la suspensión de un vehículo empleados.

Masa suspendida	m_s	250 kg
Masa no suspendida	m_u	25 kg
Rigidez del muelle	k_1	1600 N/m
Rigidez del amortiguador	C_s	980 Ns/m
Rigidez del neumático	k_2	160000 N/m

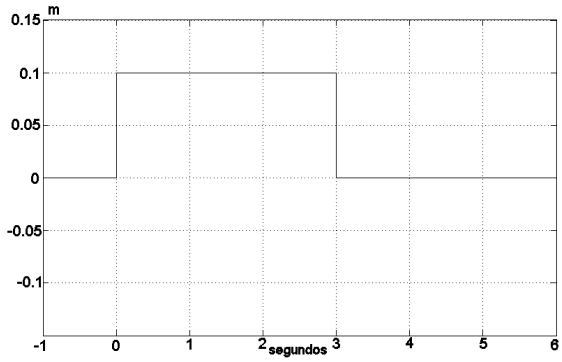

Fig. 5.a. Perfil de un terreno regular.

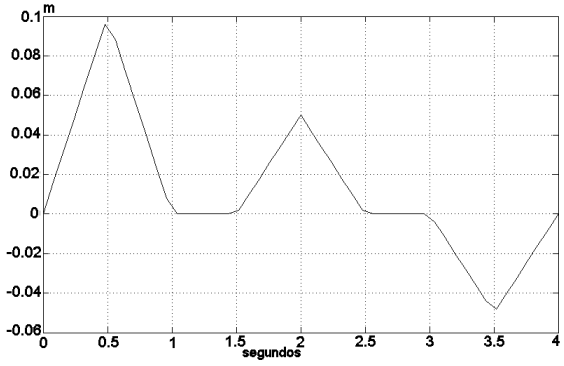

Fig. 5.b. Perfil de un terreno irregular.

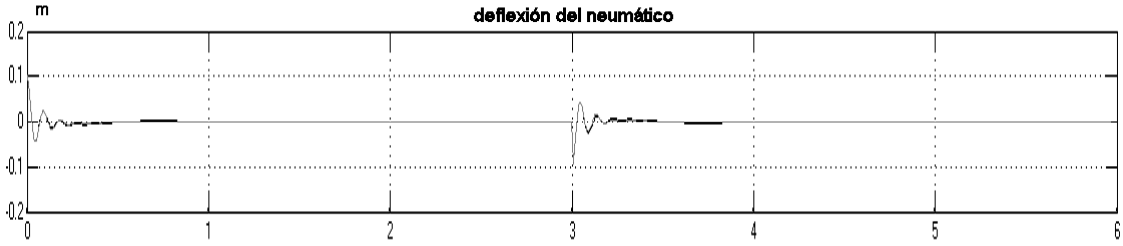

Fig. 6. Respuestas de la deflexión del neumático con suspensión pasiva (línea negra) y suspensión semi-activa (línea gris) en terreno regular.

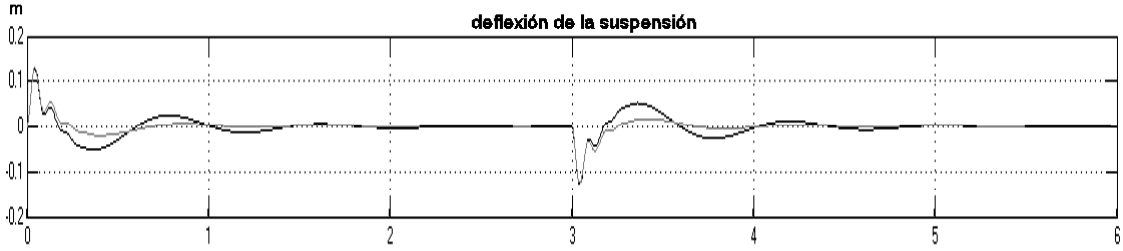

Fig. 7. Respuestas de la deflexión de la suspensión con suspensión pasiva (línea negra) y suspensión semi-activa (línea gris) en terreno regular.

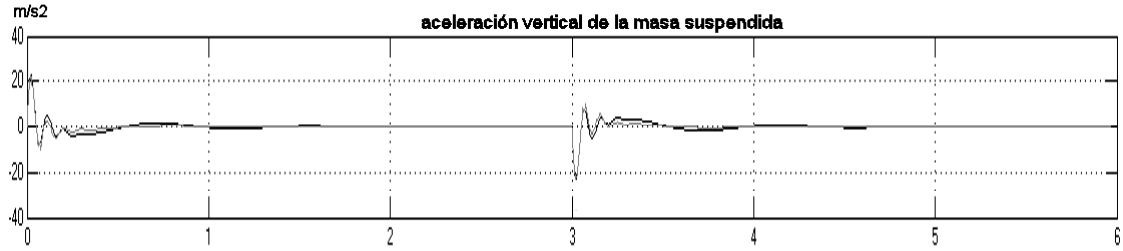

Fig. 8. Respuestas de la aceleración vertical de la masa suspendida con suspensión pasiva (línea negra) y suspensión semi-activa (línea gris) en terreno regular.

tienden a cero más rápidamente que en el sistema de suspensión pasiva. La respuesta de la deflexión del neumático en el sistema de suspensión semi-activa es prácticamente igual que en el sistema de suspensión pasiva. En este caso, se mejora el confort de los pasajeros. Los parámetros utilizados durante el aprendizaje son $\beta = 50$, $\mu = 0.3$, $\sigma_{rbf} = 0.1$, $a_{\min} = 0.2$ y $\sigma = 10$ y el número de nodos creados es 203.

Las Figuras 9 y 10 muestran el valor de la fuerza proporcionada por el controlador neuronal y la señal de refuerzo externa recibida por el controlador respectivamente después de 1153 ciclos de aprendizaje.

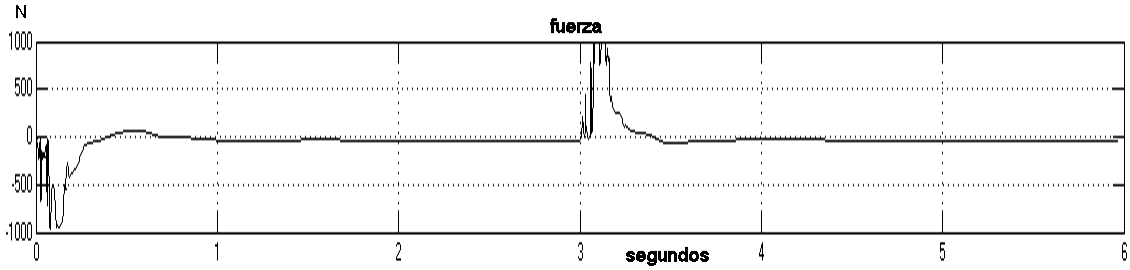

Fig. 9. Fuerza proporcionada por el controlador neuronal en terreno regular.

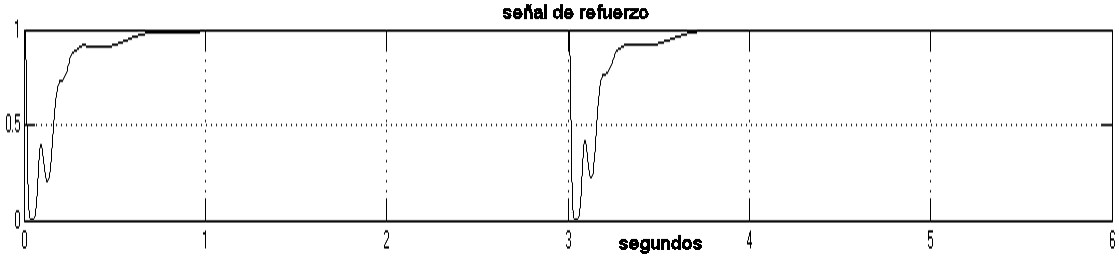

Fig. 10. Señal de refuerzo externa recibida por el controlador neuronal en terreno regular.

Las Figuras 11, 12 y 13 comparan la respuesta de un sistema de suspensión pasiva (línea negra), con la respuesta del sistema de suspensión semi-activa (línea gris) en un terreno de perfil irregular después de 527 ciclos de aprendizaje (un ciclo de aprendizaje corresponde a un periodo de 4 segundos). En este caso, el controlador neuronal mejora tanto el confort de los pasajeros y el contacto neumático-suelo. Los parámetros utilizados durante el aprendizaje son $\beta = 50$, $\mu = 0.3$, $\sigma_{rbf} = 0.1$, $a_{\min} = 0.2$ y $\sigma = 100$ y el número de nodos creados es 464.

Las Figuras 14 y 15 muestran el valor de la fuerza proporcionada por el controlador neuronal y la señal de refuerzo externa recibida por el controlador respectivamente después de 527 ciclos de aprendizaje.

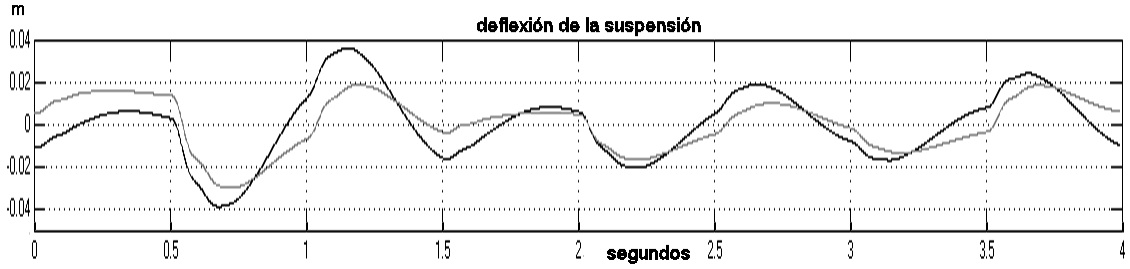

Fig. 11. Respuestas de la deflexión del neumático con suspensión pasiva (línea negra) y suspensión semi-activa (línea gris) en terreno irregular.

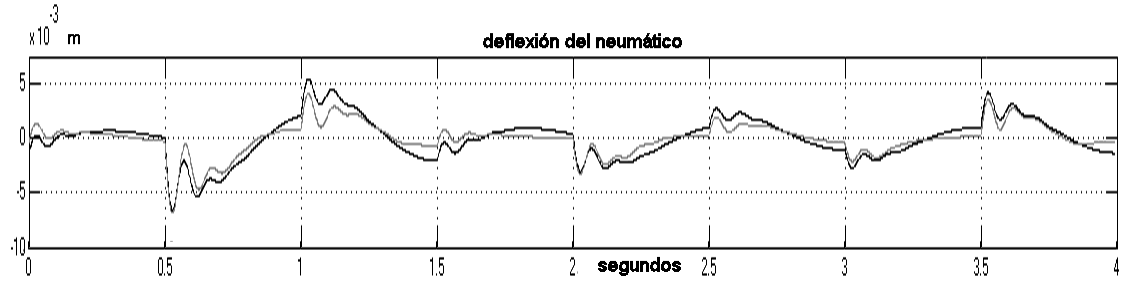

Fig. 12. . Respuestas de la deflexión de la suspensión con suspensión pasiva (línea negra) y suspensión semi-activa (línea gris) en terreno irregular.

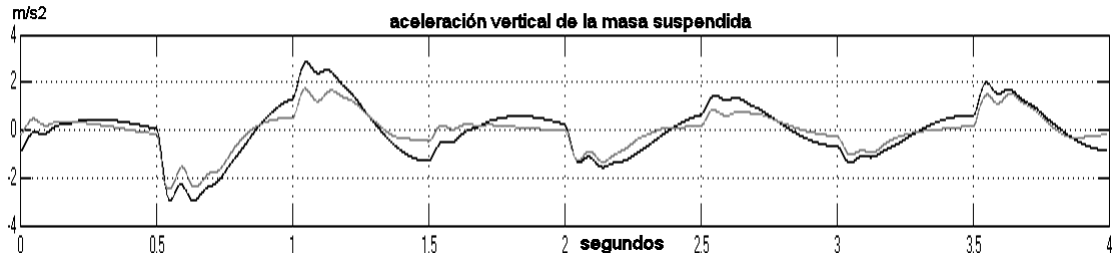

Fig. 13. Respuestas de la aceleración vertical de la masa suspendida con suspensión pasiva (línea negra) y suspensión semi-activa (línea gris) en terreno irregular.

Fig. 14. Fuerza proporcionada por el controlador neuronal en terreno irregular.

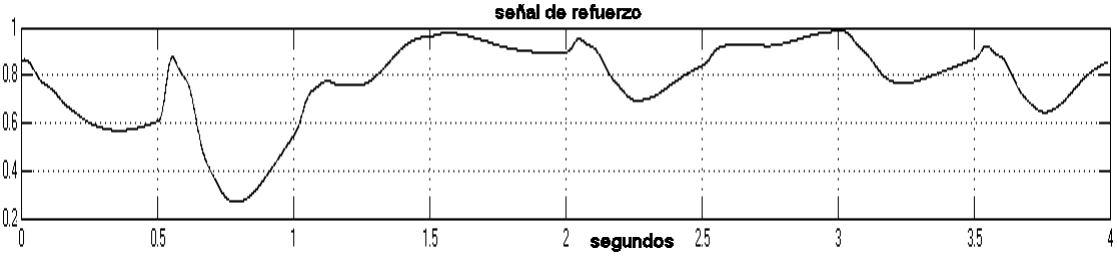

Fig. 15. Señal de refuerzo externa recibida por el controlador neuronal en terreno irregular.

Las Figuras 16 y 17 muestran las respuestas del sistema de suspensión semi-activa que ha aprendido previamente para un terreno de perfil regular como el mostrado en la Figura 5.a cuando las entradas son terrenos de perfil irregular como el mostrado en la Figura 5.b y aleatorio respectivamente. Como se puede apreciar el sistema es robusto frente a entradas diferentes a las previamente aprendidas. En ambos casos la respuesta del sistema de suspensión semi-activa (línea gris) es mejor o igual a la respuesta del sistema de suspensión pasiva (línea negra).

6. CONCLUSIONES

Este artículo presenta un controlador neuronal para el control de un sistema de suspensión semi-activa. El algoritmo de aprendizaje por refuerzo propuesto permite al sistema de suspensión semi-activa mejorar su comportamiento (confort de los pasajeros y seguridad) *on-line* tanto en terrenos regulares como irregulares. La arquitectura de red neuronal utilizada para la implementación del algoritmo de aprendizaje por refuerzo trabaja tanto con entradas y salidas continuas como con señales de refuerzo continuas. Los nodos de la capa de entrada son creados dinámicamente. Sólo son tenidas en cuenta aquellas situaciones donde el sistema ha explorado reduciéndose con ello el tamaño del espacio de entrada. Otras ventajas que presenta este algoritmo son, por una parte, que no es necesario estimar una recompensa esperada debido a que el controlador neuronal recibe una señal de refuerzo cada vez que realiza una acción y, por otra parte,

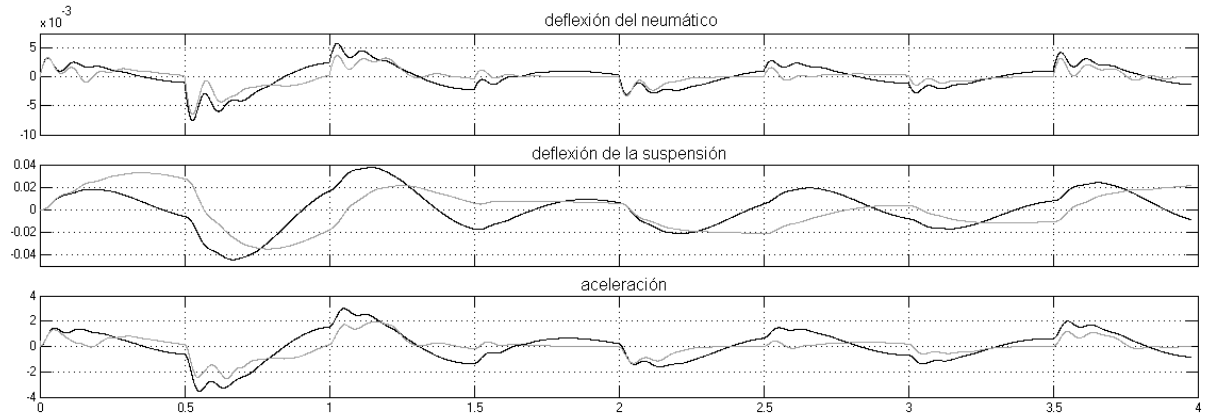

Fig. 16. Respuesta del sistema de suspensión semi-activa (línea gris), que ha aprendido en terreno regular, en terreno irregular con perfil como el mostrado en la Figura 5.b. La líneas negras representan las respuesta del sistema pasivo frente al mismo perfil de terreno.

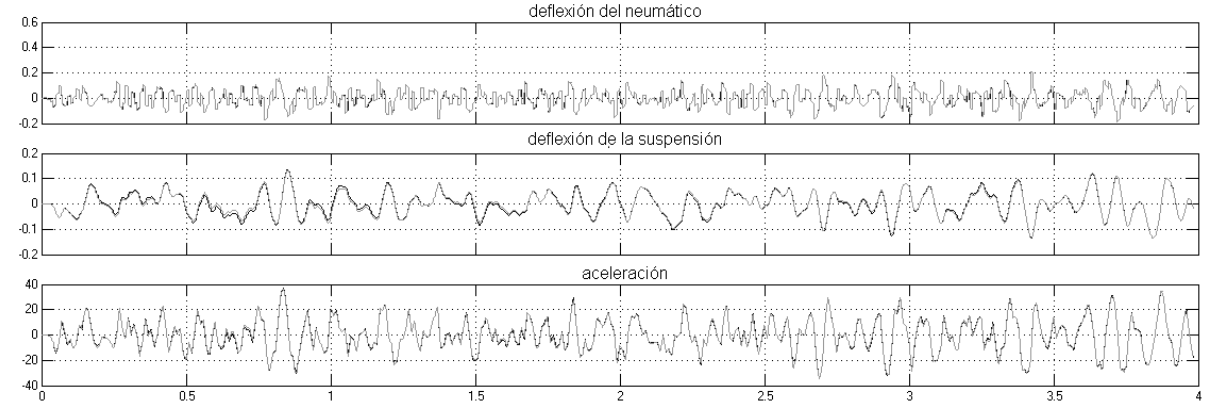

Fig. 17. Respuesta del sistema de suspensión semi-activa (línea gris), que ha aprendido en terreno regular, en terreno irregular aleatorio. La líneas negras representan las respuesta del sistema pasivo frente al mismo perfil de terreno.

que el controlador aprende *on-line*, por lo que puede adaptarse a cambios producidos en el entorno o en el propio sistema de suspensión.

De los resultados obtenidos se puede concluir además que:

- Cuánto mayor sea el número de entradas, mayor dificultad tendrá el sistema en aprender (se sugiere utilizar menos de 4 entradas).
- La elección de la señal de refuerzo externa como de los parámetros de aprendizaje (β, μ, σ, etc.) influyen de manera importante en el aprendizaje del sistema.
- No existe una regla precisa para elegir dichos parámetros de aprendizaje sino que es más bien la experiencia y el estudio de la respuesta del sistema durante el aprendizaje la que permite elegirlos más adecuadamente.
- Un sistema que evoluciona con muchos cambios y rápidamente va a ser muy difícil que aprenda con este algoritmo de aprendizaje por refuerzo.

Los resultados de simulación obtenidos muestran la efectividad del algoritmo propuesto y que el sistema aprende más rápidamente en terrenos irregulares que en terrenos regulares.

BIBLIOGRAFÍA

[1] D. S. Joe, N. Al-Holou, "Development and evaluation of fuzzy logic controller for vehicle suspension systems", in The 1995 IEEE South-eastern Symposium on System Theory, 295-299 (1995).

[2] J. Alberdi, "Amortiguadores y suspensión", Manuales de Automoción, Tecnun, www.tecnun.es/automocion (2004).

[3] N. Barbieri, "The optimal active suspension systems for an off-road vehicle", Int. Journal of Vehicle Design 16, **2/3,** 219-228 (1995).

[4] N. Yagiz, I. Yuksek, "Sliding mode control of active suspensions for a full vehicle model", Int. Journal of Vehicle Design 26, **2/3**, 264-276 (2001).

[5] T. Yoshimura, K. Watanabe, "Active suspension of a full car model using fuzzy reasoning based on single input rule modules with dynamic absorbers", Int. Journal of Vehicle Design 31, **1**, 22-39 (2003).

[6] M. Kurimoto, T. Yoshimura, "Active suspension of passenger cars using sliding mode controllers (based on reduced models)", Int. Journal of Vehicle Design 19, **4**, 402-414 (1998).

[7] Q. Li, T. Yoshimura, J. Hino, "Active suspension with preview of large-sized buses using fuzzy reasoning", Int. Journal of Vehicle Design 19, **2,** 187-198 (1998).

[8] Y. Y. Nazaruddin, M. Yamakita, "Neuro-fuzzy based modelling of vehicle suspension system", in Proceedings of the 1999 IEEE International Conference on Control Applications, 1490-1495 (1999).

[9] A. Soliman, D. Crolla, "Preview control for a semi-active suspension system", Int. Journal of Vehicle Design 17, **4,** 384-397 (1996).

[10] M. Rao, V. Prahlad, "A tuneable fuzzy logic controller for vehicle-active suspension systems", Fuzzy Sets and Systems 85, 11-21 (1997).

[11] S. Hwang, S. Heo, K. Park, "Design and evaluation of semi-active suspension control algorithms using hardware-in-the-loop simulations", Int. Journal of Vehicle Design 19, **4**, 540-551 (1998).

[12] S.-S. Han, S.-B. Choi, "Control performance of an electrorheological suspension system considering actuator time constant", Int. Journal of Vehicle Design 29, **3,** 226-242 (2002).

[13] C. F. Nicolas, J. Landaluze, E. Castrillo, M. Gaston, R. Reyerot, "Application of fuzzy logic control to the design of semi-active suspension systems", in IEEE Transaction on Fuzzy Systems, 987-993 (1997).

[14] S. G. Foda, "Neuro-fuzzy control of a semi-active car suspension system", in IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM'01), 686-689 (2001).

[15] N. Al-Holou, D. S. Joo, "Fuzzy logic based suspension system", in IEEE 1994 South-eastern Symposium on System Theory, 172-176 (1994).

[16] N. Al-Holou, T. Lahdhiri, D. S. Joo, J. Weaver, F. Al-Abbas, "Sliding mode neural network inference fuzzy logic control for active suspension systems", IEEE Transactions on Fuzzy Systems 10, **2** 234-246 (2002).

[17] T. Yoshimura, K. Nakaminami, M. Kurimoto, J. Hino, "Active suspension of passenger cars using linear and fuzzy-logic controls", Control Engineering Practice 7, 41-47 (1999).

[18] T. Yoshimura, H. Kubota, K. Takei, M. Kurimoto, J. Hino, "Construction of an active suspension system of a quarter car model using fuzzy reasoning based on single input rule modules", Int. Journal of Vehicle Design 23, **3/4**, 297-306 (2000).

[19] F. J. D'Amato, D. E. Viassolo, "Fuzzy control for active suspensions", Mechatronics 10, 897-920 (2000).

[20] W. Rattasiri, S. K. Halgamuge, "Computationally advantageous and stable hierarchical fuzzy systems for active suspension", IEEE Transactions on Industrial Electronics 50, **1,** 48-61 (2003).

[21] I. Esat, "Genetic algorithm-based optimization of a vehicle suspension system", Int. Journal of Vehicle Design 21, **2/3**, 148-160 (1999).

[22] Y. Watanabe, R. Sharp, "Neural network learning control of automotive active suspension systems", Int. Journal of Vehicle Design 21, **2/3**, 124-147 (1999).

[23] R. Venugopal, M. Beine, A. Ruekgauer, "Real-time simulation of adaptive suspension control using dspace control development tools", Int. Journal of Vehicle Design 29, **1/2**, 128-138 (2002).

[24] T. Mitchell, *Machine Learning*, McGraw Hill (1997).

[25] L. P. Kaelbling, M. L. Littman, A. W. Moore, "Reinforcement learning: A survey", Artificial Intelligent Research **4,** 237-285 (1996).

[26] M. N. Howell, G. P. Frost, T. J. Gordon, Q. H. Wu, "Continuous action reinforcement learning applied to vehicle suspension control, Mechatronics 7, **3,** 263-276 (1997).

[27] S.-J. Huang, W.-C. Lin, "Adaptive fuzzy controller with sliding surface for vehicle suspension control", IEEE Transactions on Fuzzy Systems 11, **4**, 550-589 (2003).

[28] Yueh-Jaw, "Toward better ride performance of vehicle suspension", in IEEE Conference on System, Man and Cybernetics, 1470-1475 (1992).

[29] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, The MIT Press, (1998).

[30] S. Singh, T. Jaakkola, M. Littman, C. Szepesvari, "Convergence results for single-step on-policy reinforcement-learning algorithms", Machine Learning 38, **3**, 287-308 (2000).

[31] V. Gullapalli, "Reinforcement learning and its application to control", Ph.D. thesis, Birla Institute of Technology and Science. University of Massachusetts (1992).

[32] R. S. Sutton, "Learning to predict by the method of temporal differences, Machine Learning" 3, **1**, 9-44 (1988).

[33] A. G. Barto, R. S. Sutton, C. W. Anderson, "Neurolike elements that can solve difficult learning control problems", in IEEE Transactions on Systems, Man and Cybernetics, **13**, 835-846 (1983).

[34] C. J. H. Watkins, P. Dayan, Q'learning, Machine Learning 8, **3**, 279-292 (1988).

[35] D. Rumelhart, G. Hinton, R. Williams, "Learning representations by backpropagation errors", Nature 323, 533-536 (1986).

[36] C. W. Anderson, "Strategy learning with multi-layer connectionist representations", in Proceedings of the Fourth International Workshop on Machine Learning, 103-114 (1987).

[37] T. C.-K. Hui, C.-K. Tham, "Adaptive provisioning of differentiated services networks based on reinforcement learning", IEEE Transactions on Systems, Man and Cybernetics, Part C 33, **4**, 492-501 (2003).

[38] P. Maes, R. A. Brooks, "Learning to coordinate behaviours", in Proceedings, AAAI-90, 796-802 (1990).

[39] J. A. Bagnell, J. G. Schneider, "Autonomous helicopter control using reinforcement learning policy search methods", in IEEE International Conference on Robotics and Automation, **2**, 1615-1620 (2001).

[40] C. Zhou, Q. Meng, "Dynamic balance of a biped robot using reinforcement learning agents", Fuzzy Sets and Systems 134, **1**, 169-187 (2003).

[41] T. P. I. Ahamed, P. S. N. Rao, P. S. Satry, "A reinforcement learning approach to automatic generation control", Electric Power Systems Research 63, **1**, 9-26 (2002).

[42] N. Krodel, K.-D. Kuhner, "Pattern matching as the nucleus for either autonomous driving or driver assistance systems", in Proceedings of the IEEE Intelligent Vehicle Symposium, 135-140 (2002).

[43] M. C. Choy, D. Srinivasan, R. L. Cheu, "Cooperative, hybrid agent architecture for real-time tra±c signal control", IEEE Transactions on systems, man, and cybernetics: PART A. 33, **5**, 597-607 (2003).

[44] I. O. Bucak, M. A. Zohdy, "Application of reinforcement learning to dextereous robot control", in Proceedings of the 1998 American Control Conference. ACC'98, **3**, USA, 1405-1409 (1998).

[45] D. F. Hougen, M. Gini, J. Slagle, "Rapid unsupervised connectionist learning for backing a robot with two trailers", in IEEE International Conference on Robotics and Automation., 2950-2955 (1997).

[46] F. Fernández, D. Borrajo, VQQL. "Applying Vector Quantization to Reinforcement Learning", Lecture Notes in Computer Science, 292-303 (2000).

[47] S. Yamada, M. Nakashima, S. Shiono, "Reinforcement learning to train a cooperative network with both discrete and continuous output neurons", IEEE Transactions on Neural Network 9, **6**, 1502-1508 (1998).

[48] A. J. Smith, "Applications of the self-organising map to reinforcement learning", Neural Network 15, **8-9**, 1107-1124 (2002).

[49] M. J. L. Boada, R. Barber, M. A. Salichs, "Visual approach skill for a mobile robot using learning and fusion of simple skills", Robotics and Autonomous Systems, 38, 157-170 (2002).

[50] M. J. L. Boada, V. Egido, R. Barber, M. A. Salichs, "Continuous reinforcement learning algorithm for skills learning in an autonomous mobile robot", in IEEE International Conference on Industrial Electronics Society, IECON'02, (2002).

[51] A. S. Soembagijo, H. V. Brussel, "Robot visual tracking control using neural networks", Intelligent Autonomous Systems, 562-68 (1995).

CONTINUOUS REINFORCEMENT LEARNING ALGORITHM FOR SEMI-ACTIVE SUSPENSION CONTROL

Abstract – This paper presents a reinforcement learning algorithm using neural networks which allows a vehicle with semi-active suspension to improve continuously the ride comfort and the tyre/ground contact both flat terrain and even terrain. The implemented neural network architecture works with continuous input and output spaces, has a good resistance to forget previously learned actions and learns quickly. Other advantages this algorithm presents are that on one hand, it is not necessary to estimate an expected reward because the controller receives a real continuous reinforcement each time it performs an action and, on the other hand, the system learns on-line, so that the system can adapt to changes produced in the environment. The proposed neuro controller has been studied using a quarter vehicle model. The results show the effectiveness of our algorithm.