UNIVERSIDAD CARLOS III DE MADRID

DEPARTAMENTO DE TECNOLOGÍAS DE LAS COMUNICACIONES

TESIS DOCTORAL

TÉCNICAS DE IGUALACIÓN NO LINEAL VÍA ESTRUCTURAS EN ESCALERA

AUTOR: Manel Martínez Ramón
DIRECTOR: Prof. Dr. Aníbal Ramón Figueiras Vidal

JUNIO, 1999

AGRADECIMIENTOS

Siempre tendré presente la memoria del Profesor Jaume Aranda, Doctor en Ciencias Físicas por la Universitat de Barcelona. Gracias a todo lo que aprendí de él y al empeño que siempre puso en ello, he podido realizar la aspiración de elaborar una tesis doctoral. Nunca, ni una sola vez, pude darle las gracias, porque se fue demasiado pronto; antes de que me diera cuenta de todo lo que hizo por mí.

Quiero agradecer todo el esfuerzo que en la dirección y corrección de este trabajo ha puesto mi director, el Profesor Aníbal Ramón Figueiras Vidal. Siendo esta tesis tan modesta como ustedes quieran, creo que reconocerán que es, gracias él, digna, y fruto de un trabajo intenso y excelentemente dirigido. Además, ha resultado el trabajo más grato que he hecho hasta ahora en toda mi vida. Reciba Vd. mi más sincero agradecimiento por todo lo dicho y por la confianza que ha depositado en mí.

Gracias también a los Profesores Miguel Ángel Lagunas, Gregori Vázquez y Juan Antonio Fernández Rubio, de la Universitat Politècnica de Catalunya. Ellos me abrieron la puerta al mundo universitario, me animaron a trabajar en una tesis y, además el Dr. Lagunas me presentó al Dr. Figueiras.

También debo mencionar al Dr. Ramón López de Arenosa, quien, trabajando yo a su cargo en la CICYT, siempre dio prioridad a mi tesis por encima de algunas de mis obligaciones. Eso también va para el Dr. Eduardo Artal.

La elaboración de esta tesis ha pasado por una serie de incidencias, la última de las cuales fue la pérdida irreparable de todos los datos y programas, elaborados durante los últimos tres años, el día 21 de diciembre de 1998. Esto viene a cuento porque enseguida pude observar la solidaridad que gastan mis compañeros de trabajo, a quienes no cito porque son demasiado numerosos como para hacerlo y, además, corro el peligro de dejarme a alguien fuera, lo que sería imperdonable. Sin embargo, permítanme que sí cite a Greg y Jeri Heileman, a quienes también incluyo aquí. Gracias a todos ustedes, pude rehacer (que no recuperar) los datos y continuar con mi tesis. Hay otras cosas que agradecerles, pero ahora esto me parece lo más importante. Quizá les parecerá que no hicieron nada, pero cada uno dio su brazo a torcer en el momento adecuado y en las formas más variadas. Lo que recordaré siempre con mucho cariño.

Gracias, por otra parte, al Director de Doctorado, Dr. Fernando Díaz de María, que hizo un excelente trabajo, por encima de sus obligaciones, en la gestión de esta tesis.

No puedo dejar de mencionar a mis compañeros de la Agrupación Musical Vox Aurea. Estos tres últimos años han sido menos duros al haber podido compartir con este grupo mi otra pasión: la música.

Tampoco sería justo no mencionar a mis amigos del grupito de Sant Pere de les Puelles. Ha sido muy grato el apoyo que me han prestado.

Eso también va por Javier Francisco y su esposa Carmen Navarro. Más que amigos, son mis hermanos no carnales desde ya tiempo inmemorial, y no puedo menos que acordarme de ellos en este momento tan importante para mí.

Debo agradecer al Profesor Jaume Riba i Sagarra, todos los consejos que me ha dado y todo lo demás que usted ya sabe. En parte, me metí en esto siguiendo sus pasos. Tiene Vd. toda mi admiración.

Gracias a Marta, mi esposa, que se leyó toda esta tesis y encontró ciento setenta y tres gazapos de diversa índole. Pero no sólo tengo que agradecerle esto, sino todos los fines de semana, vacaciones, etc., que ha sacrificado por mi culpa. Y todo el mal humor que ha tenido que soportar. Y muchas más cosas. *Gràcies, Marteta.*

Y Gracias a Dios por haber puesto delante de mí a todas estas personas.

A la meva esposa Marta,
que es el més important de ma vida.
"Tout ce que j'ai pu écrire
je l'ai puisé
à l'encre de tes yeux"

Als meus germans Katy i Mario.

A ma mare i a mon pare.

RESUMEN

Los esquemas de igualación no lineal permiten aumentar significativamente la velocidad de transmisión de datos en canales dispersivos, y particularmente, en los que además hay no lineales.

Sus inconvenientes más importantes son dos: su alta complejidad y su baja velocidad de convergencia. Los esquemas no lineales necesitan una gran potencia de cálculo para poder llevar a cabo el elevado número de operaciones necesarias por dato adquirido en el receptor. Por otra parte, algunos esquemas no lineales como el perceptrón multicapa necesitan de un gran número de datos para poder llegar a la convergencia y, además, corren el riesgo de caer en mínimos locales de convergencia, lo que los hace poco prácticos en comunicaciones. Otros esquemas, como los filtros de Volterra, son menos propensos (por su arquitectura) a caer en mínimos locales, pero son difíciles de ajustar para que converjan.

En algunas situaciones no es necesario el uso de esquemas no lineales para alcanzar probabilidades de error razonablemente bajas. Muchas veces se puede considerar que el canal es localmente lineal. Si se utilizan esquemas basados en funciones lineales que actúen localmente e independientes unos de otros, se consiguen resultados satisfactorios. Incluso cuando el canal presenta no linealidades, es posible utilizar esquemas lineales que, actuando en paralelo, se especialicen en zonas diferentes del espacio.

En esta Tesis se examinará un esquema de este tipo, que es el esquema en escalera. Este esquema se basa en una cadena de decisiones bit a bit a partir de las cuales se lleva a cabo la decisión multinivel.

El procedimiento bit a bit del esquema en escalera permite introducir algoritmos de gradiente basados en objetivos binarios, lo que permite equilibrios en compromisos en prestaciones frente a velocidad. Sin embargo, algunos objetivos, como la Entropía Relativa, sólo convergen bajo ciertas condiciones iniciales, lo que dificulta el funcionamiento del sistema. Aquí se proponen maneras sencillas de facilitar la convergencia sin necesidad de introducir restricciones en el sistema.

El esquema en escalera facilita la inserción de algoritmos no lineales sólo en las zonas del espacio que lo necesiten, dejando las otras para clasificación lineal.

Por otro lado, se han desarrollado sistemas de clasificación inspirados en la Máquina de Vectores de Soporte, que ha dado muy buenos resultados en problemas de clasificación, pero que no es adaptativa.

En esta Tesis desarrollamos variantes de este algoritmo basados en métodos de selección de muestras, que han proporcionado buenos resultados en velocidad de convergencia y capacidad de seguimiento de canales no estacionarios.

ABSTRACT

Nonlinear equalization schemes significantly enable data transmission speed to be increased in dispersive channels, and particularly in those which are also nonlinear.

Two are its most important drawbacks: its high complexity and its low convergence rate. Nonlinear schemes need high computational power so as to be able to carry out the big number of necessary computations per sample. On the other hand, some nonlinear schemes such as the multilayer perceptron not only need a large amount of data in order to reach convergence, but they also run the risk of falling down to local minima for convergence, which makes them be little practical in communications. Other schemes, such as Volterra filters, are less prone to fall down to local minima thanks because of their architecture, but they do not easily adjust for reaching convergence.

In some situations, the use of nonlinear schemes for reaching reasonably low error probabilities is not necessary. In many cases one can consider that the channel is localy linear. When schemes based upon linear functions operating localy and independently from one another are used, one obtains satisfactory results. Even if the channel has nonlinearities, it is possible to use linear schemes which, operating in parallel, specialize in different areas of the space.

In this Thesis this kind of scheme, called staircare scheme, will be analysed. It is based on a bit by bit decisions chain from which one carries out the decision in a communications multilevel.

The bit by bit staircase scheme procedure allows gradient algorithms based upon binary objectives to be introduced, which enables balancing commitments performance and speed. Nevertheless, some objectives, such as Relative Entropy, only converge under certain initial conditions which hinder the running of the system. Here they suggest easy ways of making convergence easy without the need of introducing any restrictions in the system.

The staircase scheme makes nonlinear algorithm insertion easier only in those areas of the space where it is needed, leaving the remaining nonlinear algorithms to linear classification.

On the other hand, classification systems inspired by the Support Vectors Machine have shown good results in classification problems tough they are not adaptative.

In this thesis we will develop variants of this algorithm based on sample selection methods which have fournished good results in speed convergence as well as in their nonstationary channel tracking abilty.

ÍNDICE

1. LA DISTORSIÓN EN COMUNICACIONES DIGITALES

1.1. INTRODUCCIÓN

Está demostrado que la transmisión digital de señales ofrece suficientes ventajas como para implantarse en todos los sistemas de comunicación excepto en aquellas pocas aplicaciones en las que su sencillez no justifique el incremento del coste que significa la tecnología digital. En los casos en que los datos son inherentemente digitales (ordenadores), la transmisión digital es la única alternativa: en otros, la tecnología analógica se ha utilizado durante decenios para la transmisión de señales debido a la imposibilidad técnica o económica de digitalizar la información; sin embargo, las ventajas de la tecnología digital no se limitan a la transmisión de señales. El tratamiento de la información (en banda base) es mucho más versátil si los procesos se llevan a cabo digitalmente: el filtrado (lineal) adaptativo, el cifrado, la compresión de información, la codificación y el mero almacenamiento justifican por sí solos el uso de la digitalización de señales. Además, aún teniendo en cuenta las limitaciones de la tecnología para realizar circuitos electrónicos digitales de alta frecuencia (limitaciones que cada vez son menores) estos circuitos ofrecerán mucha más versatilidad que sus equivalentes analógicos por una razón muy sencilla: los digitales son programables y, por lo tanto, cada aplicación específica necesita de una menor especialización.

Por todo ello, la tendencia actual hace que la señal se trate de forma analógica solamente en la transducción de la energía que transporta la información y en las etapas muy tempranas del acondicionamiento de ésta. Por ejemplo, en los últimos años ha tomado forma la idea de "Radio Software", según la cual un receptor de radiofrecuencia se debe diseñar, a partir de la etapa de frecuencia intermedia, íntegramente con uno o varios procesadores digitales de señales (DSP). Estos procesadores presentan una arquitectura y una filosofía de programación orientadas al tratamiento de la señal en el tiempo y en la frecuencia. Alguien ha definido estos dispositivos como máquinas especializadas en sumar y multiplicar a toda velocidad. La idea de "Radio Software" ha sido ya objeto de atención por parte de la industria, de tal manera que han sido presentados algunos prototipos de teléfono celular basados en tres DSP trabajando conjuntamente. Este interés también se extiende a la radiodifusión en su actual transición a la tecnología digital.

Este capítulo presenta la problemática de las comunicaciones digitales a través de canales paso banda y las soluciones clásicas basadas en el procesamiento discreto de señales.

1.2. FUENTES DE DISTORSIÓN

El canal de comunicación es el medio físico, natural o fabricado, que conecta los puntos entre los que se pretende establecer la comunicación a distancia: es el elemento de la cadena de comunicación que no puede modificarse. Cada canal de comunicación presenta unas características que determinarán en cada caso las ventajas e inconvenientes en la comunicación. El diseño de la cadena de comunicación depende, por tanto, del canal que se utilice para la transmisión. La transmisión digital permite luchar de una manera más eficiente contra las dificultades que plantea el uso de un canal de comunicaciones determinado. Todo indica que, en el futuro, los canales de comunicaciones a explotar serán las fibras ópticas y los canales de radio en la banda de las microondas, aunque existe un interés por los canales de par telefónico y las líneas coaxiales dado que se seguirán utilizando durante mucho tiempo, así como por los canales radio en bandas de frecuencias inferiores a las de microondas.

Entre las restricciones que impone un determinado canal de comunicación, dos son muy importantes: la primera es el ancho de banda del canal, que imposibilitará al sistema transmitir señales por encima de una cierta velocidad. La segunda es la potencia máxima que el transmisor puede emitir: tal potencia máxima estará limitada por la etapa de potencia de aquél, cuya linealidad estará comprometida con el margen dinámico de la salida. Es importante conocer

bien estas dos restricciones porque, junto con el ruido e interferencias que todas las etapas de la cadena de transmisión introducirán, limitarán la velocidad máxima de transmisión. El aumento de las demandas de acceso y calidad de las comunicaciones exige velocidades de transmisión de datos cada vez mayores: la telefonía no se limita hoy a la transmisión de voz, sino que se ha generalizado; por ejemplo, el uso de Internet ha precipitado el nacimiento de toda una serie de servicios en línea que demandan cada vez una mayor velocidad de transmisión.

Para aumentar la velocidad de transmisión en un determinado canal de comunicaciones paso banda es necesario aumentar el número de símbolos del alfabeto utilizado para la transmisión. En general, los símbolos transmitidos son la amplitud de un conjunto de funciones base ortogonales, que son las que se transmitirán a través del canal. La representación en el plano R^2 del conjunto de símbolos transmitidos se denomina constelación. Aquellas constelaciones que aprovechan más el ancho de banda del canal de comunicaciones son las de tipo modulación de amplitud en cuadratura (QAM), en la que las amplitudes posibles de cada una de sus dos funciones de la base ortogonal están equiespaciadas; sin embargo, cuanto más densa sea la constelación QAM mayor relación de señal a ruido y mayor linealidad del canal será requerida. Las modulaciones de desplazamiento de fase (PSK) se construyen de manera que la combinación lineal de los elementos de la base ortogonal tenga módulo constante: la información viaja en la fase de la señal. Al tener amplitud constante, la modulación presenta mayor robustez frente a la saturación en amplitud pero, como los símbolos están más cerca, el ruido aditivo degradará más la calidad de la información recibida. Cuando se transmite una constelación en la que cada símbolo es ortogonal a los demás, la transmisnión se llama QPSK. La Figura 1-1 muestra una constelación 16-PSK y una constelación 16-QAM, donde los dos ejes representan las

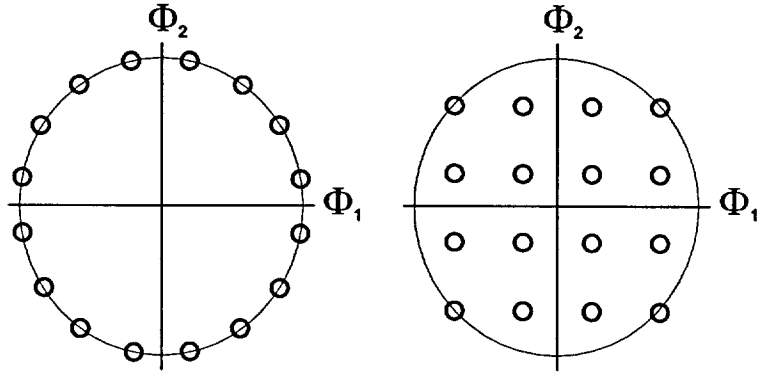


Figura 1-1. Constelación M-PSK (izquierda) y M-QAM para M=16 símbolos. La constelación de la derecha es muy robusta frente a canales que presenten compresión de amplitud. La constelación de la derecha es más eficiente en cuanto a velocidad de transmisión por cuanto los símbolos están más separados y por tanto más protegidos frente a ruido, aunque es más vulnerable a la compresión. A menudo no se utilizan los símbolos más externos de esta constelación para reducir la pérdida de calidad debida a la compresión.

amplitudes posibles de las funciones ortogonales Φ_1 y Φ_2. La constelación QAM se verá más afectada por la compresión en las partes más externas de la constelación: sin embargo, los símbolos están más juntos unos de otros en la constelación PSK que en la constelación QAM. Por esto, a igualdad de probabilidad de error, la velocidad de transmisión utilizando constelación QAM será mayor que utilizando constelación PSK

La respuesta en frecuencia del canal producirá distorsión lineal en los datos transmitidos. En algunos canales, como es el caso de la telefonía celular o la telefonía en recintos cerrados ("indoor"), se produce propagación multicamino. Ésta consiste en la propagación a través de diferentes caminos que hacen que la señal llegue al receptor con diferentes retardos, fases y amplitudes: las consecuencias de este fenómeno son a) la interferencia entre símbolos, cuando el retardo es significativo en términos de la velocidad de símbolo y, b) el desvanecimiento por interferencia destructiva de la onda recibida. El desvanecimiento por interferencia destructiva afectará más cuanto mayor sea el ancho de banda de la señal a transmitir y la interferencia intersimbólica afectará más cuanto más densa sea la constelación: la misma propagación multicamino que prácticamente no afectará a una transmisión BPSK afectará a una 256 QAM con la misma velocidad de símbolo hasta el punto de no poder efectuar correctamente la detección. Si se quiere aumentar la velocidad de transmisión, es necesario filtrar la señal en recepción en un proceso que se ha dado en llamar igualación. Además, si el canal es variante, como claramente lo es en el caso de las comunicaciones móviles, el proceso de igualación debe ser adaptativo o capaz de seguir las variaciones del canal.

Shannon estableció el límite superior del régimen binario (o número de bits transmitidos por unidad de tiempo) de un canal en relación con el ancho de banda B, la potencia media recibida P y la potencia de ruido en recepción N_0. Este límite superior, en bits por segundo, tiene la expresión [Haykin, 1988] (tercer teorema de Shannon o teorema de Shannon-Hartley):

$$C = B \log_2 \left(1 + \frac{P}{N_0 B} \right) \tag{1-1}$$

El régimen binario máximo de un canal está limitado por el ancho de banda del canal, la potencia recibida e, inversamente, por el ruido en la recepción: si se quiere aumentar la velocidad de transmisión para un determinado ancho de banda del canal debe hacerse más densa la constelación utilizada; pero como los símbolos disminuirán su distancia, debe aumentar la relación de señal a ruido (P/N_0) de la transmisión, lo que equivale, casi siempre, a aumentar la potencia de la señal emitida. Pero la potencia de la señal emitida está limitada, en particular,

por la linealidad del amplificador de potencia del transmisor: aparece un compromiso entre régimen binario y linealidad de transmisión.

1.2.1. Distorsión lineal

Los canales físicos por los que se desea enviar energía que transporte información tienen un ancho de banda determinado por su naturaleza. Así, un par telefónico trenzado o una línea coaxial pueden ser modelados mediante secciones diferenciales, cada una de las cuales contiene una inductancia en serie y una capacidad en paralelo, más dos resistencias que modelan respectivamente las pérdidas en los conductores y en el material dieléctrico que los separa.

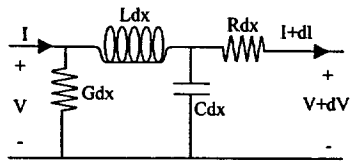


Figura 1-2. Modelo de línea de transmisión con pérdidas.

La respuesta en frecuencia de este canal será paso bajo, con una frecuencia de corte muy baja para el caso de par trenzado, de manera que lo hace apto solamente para transmisiones en la banda vocal, y mucho más elevada para el cable coaxial, que así puede transmitir energía en la banda de microondas.

La impedancia de entrada a una línea de estas características puede aproximarse por:

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \qquad (1\text{-}2)$$

Las características principales de este tipo de canales son:

- la dispersión, o fenómeno según el cual las diferentes componentes frecuenciales de la señal se propagan a velocidades decrecientes (siendo nula la velocidad de propagación a la frecuencia de corte de la línea), y
- la reflexión de la señal en las terminaciones de la línea. En microondas esto se debe a la desadaptación de las diferentes etapas, pero hay fenómenos de otras naturalezas que los producen en los cables de baja frecuencia (donde baja frecuencia significa que la longitud de la línea es mucho menor que la longitud de onda mínima). En telefonía, las bobinas híbridas suelen ser las causantes de la reflexión de la señal.

Los canales de radiofrecuencia son aquéllos en los que se transmiten ondas electromagnéticas a través del espacio a frecuencias desde algunos

kilohercios (onda larga) hasta los gigahercios (microondas). Un caso de especial interés es el de la telefonía móvil, donde los estándares utilizan frecuencias del orden de 0.5 GHz (telefonía analógica) y 0.9 GHz (digital). Los nuevos estándares digitales utilizan frecuencias del orden de 2 GHz. Estos canales no presentan distorsión lineal apreciable en las bandas de frecuencia utilizadas, excepto por la respuesta al impulso de las antenas y otros elementos del emisor y el receptor. Para frecuencias más allá de los 100 MHz se puede afirmar que el canal tiene una atenuación que disminuye con la frecuencia según la ecuación

$$L = \left(\frac{4\pi fr}{c}\right)^2 \tag{1-3}$$

siendo f la frecuencia, r la distancia entre el emisor y el receptor y c la velocidad de la luz en el vacío. De modo que, por ejemplo, si se tiene una transmisión con un ancho de banda de 2 MHz y una frecuencia central del orden de 1 GHz, la diferencia de atenuación entre los extremos de la banda no excede del 0,4% (0.034 dB). En otras palabras, la respuesta de los canales de radio está determinada casi exclusivamente por el filtro de salida del emisor (impuesto por la máscara de emisión o límites legales de ancho de banda de un emisor para transmitir en una banda determinada de frecuencias) y por el fenómeno de la propagación multicamino.

1.2.2. Pulsos de Nyquist

Al estar limitado el ancho de banda por el filtro de salida del emisor, la respuesta al impulso de éste producirá un ensanchamiento temporal del símbolo, lo que generará interferencia intersimbólica. Nyquist [Carlson, 1986] estableció unas condiciones en el dominio de la frecuencia para la forma de onda de los pulsos recibidos que, si se cumplen, hacen que esta interferencia intersimbólica desaparezca. Concretamente, el teorema de simetría vestigial de Nyquist afirma que para un pulso $p_\beta(t)$ que cumpla la condición

$$\begin{aligned} &P_\beta(f) = 0 \qquad |f| > B \\ &B = \frac{1}{2T} + \beta; \qquad 0 \le \beta \le \frac{1}{2T} \end{aligned} \tag{1-4}$$

el pulso p(t) construido en la siguiente forma

$$P(f) = P_\beta(f)\,\mathrm{sinc}\left(\frac{t}{T}\right) \tag{1-5}$$

cumple la condición

$$p(t) = \begin{cases} 1 & t = 0 \\ 0 & t = \pm T, \pm 2T \end{cases} \tag{1-6}$$

donde T es el periodo de símbolo, lo que asegura la ausencia de interferencia intersimbólica en los instantes de muestreo. Los pulsos que cumplen la condición anterior se denominan pulsos de Nyquist. En la práctica no se pueden cumplir estas condiciones en forma exacta, pero sí en forma muy aproximada. Sin embargo, el pulso de Nyquist no asegura la ausencia de interferencia intersimbólica en aquellos canales que presenten propagación multicamino, por lo que es preceptivo eliminarla por otros métodos.

1.2.3. Propagación multicamino

La propagación multicamino es un fenómeno producido por las condiciones físicas del canal, que normalmente serán en todo o en parte desconocidas, y además serán, en general, variantes con el tiempo. La atmósfera es no homogénea y puede que la radiación electromagnética en el camino desde el emisor al receptor encuentre regiones con diferentes índices de refracción, lo cual produce diferentes velocidades de transmisión. Esto se traduce en que la señal llegará al receptor por dos o más caminos con diferencias en el instante de llegada de los frentes de onda. Las condiciones ambientales tienen una lenta variación a lo largo del día, lo que hace que este tipo de fenómenos sea variante. También es común encontrar canales de transmisión con obstáculos tales como edificios (en el caso de telefonía móvil) o accidentes geográficos (por ejemplo, en los radioenlaces): lo que, unido a transmisiones con antenas omnidireccionales o con aperturas no muy estrechas o con lóbulos secundarios apreciables, hace que la señal llegue al receptor por diferentes caminos, con diferentes atenuaciones y fases. La propagación multicamino significa que el canal tiene una atenuación dependiente de la frecuencia, atenuación apreciable dentro del ancho de banda de la transmisión por estrecho que sea éste.

El efecto de la propagación multicamino se puede determinar utilizando el modelo equivalente paso bajo de su función de transferencia. Sirva como ejemplo un canal con propagación multicamino a través de dos caminos con distancias d_1 y d_2 y atenuaciones de propagación A_1 y A_2. Los retardos de propagación serán,

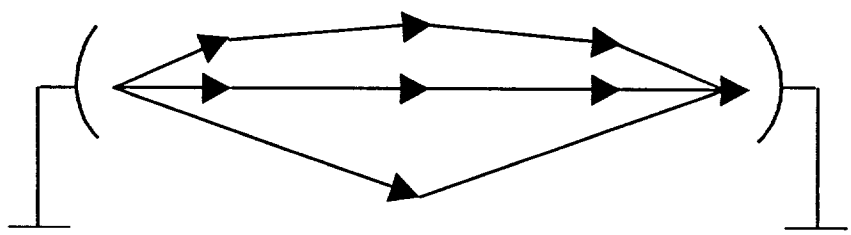
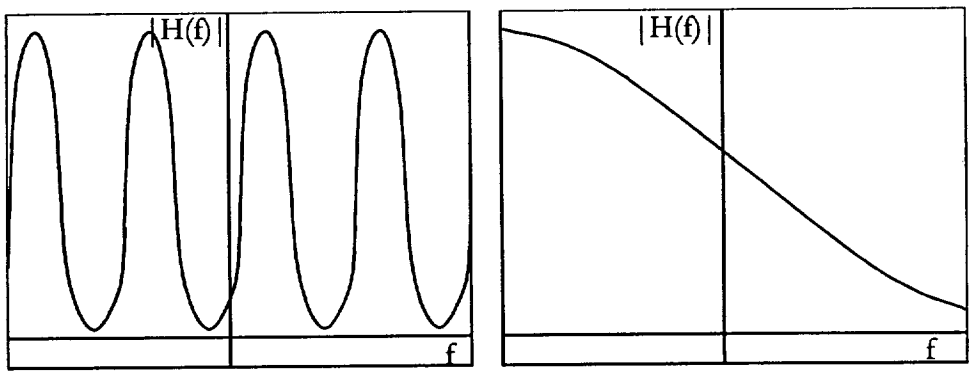


Figura 1-3. Propagación multicamino



Canal de banda ancha Canal de banda estrecha

Figura 1-4. Respuesta frecuencial de un canal de banda ancha y uno de banda estrecha con propagación por dos caminos con $A_2/A_1 < 1$

respectivamente, $\tau_1=d_1/c$ y $\tau_2=d_2/c$. Para una determinada frecuencia ω, el equivalente paso bajo correspondiente a la señal que llega a la antena receptora es, por superposición lineal:

$$A_1 e^{-j\omega\tau_1} e^{-jkd_1} + A_2 e^{-j\omega\tau_2} e^{-jkd_2} = A_1 e^{-j\omega\tau_1} e^{-jkd_1} \left(1 + \frac{A_2}{A_1} e^{-j\omega\Delta\tau} e^{-jk\Delta d} \right) \quad (1\text{-}7)$$

siendo $\Delta d=d_2 - d_1$ y $\Delta\tau=\tau_2-\tau_1$. El término entre paréntesis constituye la respuesta del canal y depende de la diferencia de frecuencias y de distancias entre los dos caminos. Obsérvese que la interferencia puede ser destructiva y anular la amplitud de la señal recibida. Pueden distinguirse dos casos. En el primero, el retardo temporal es pequeño en relación con el periodo de la señal; es decir, ω es suficientemente pequeña y, por tanto, puede decirse que $|\omega\Delta\tau| \ll \pi$. En este caso la variación de la amplitud es casi lineal con la frecuencia. En el caso en que la variación de ω sea suficientemente grande, la respuesta frecuencial presentará mínimos a lo largo de todo el ancho de banda, en los puntos en que $|\omega\Delta\tau|=\pi$. La Figura 1-4 muestra el aspecto del módulo de la respuesta en frecuencia para los dos casos. Obsérvese que esta respuesta no tiene por qué ser real, es decir, afectará tanto al módulo como a la fase de la señal recibida.

1.2.4. Interferencia intersimbólica (ISI)

Cuando la comunicación radioeléctrica se establece en un entorno tal como el urbano, pueden identificarse multitud de caminos de propagación con distancias considerablemente diferentes. Los retardos entre diferentes caminos de propagación pueden ser superiores a 10 microsegundos, de manera que para transmisiones a velocidades mayores de 100 KBaudios la interferencia intersimbólica (ISI) será importante. La respuesta al impulso de un canal con

propagación multicamino por K caminos distintos puede expresarse de la siguiente forma:

$$h(t) = \sum_{k=0}^{K-1} h_k' \delta(t - t_k) \tag{1-8}$$

siendo h_k' una cantidad compleja que representa la fase y la amplitud que introduce el camino k, y t_k el retardo total introducido por este camino. Se puede elegir $t_0=0$.

Si el canal presenta estas características, la señal de entrada al receptor estará afectada de interferencia intersimbólica ya que, por un lado, los símbolos que lleguen por diferentes caminos se solaparán, y por otro, no se podrá asegurar la condición de no ISI que, en ausencia de propagación multicamino, proporcionan los pulsos de Nyquist.

En efecto, si la señal de entrada al canal tiene la forma

$$x(t) = \sum_{m=-\infty}^{\infty} a_m p_\beta (t - mT) \tag{1-9}$$

la salida del canal será

$$\begin{aligned} y(t) &= \sum_{m=-\infty}^{\infty} a_m p_\beta (t-mT) * \sum_{k=0}^{K-1} h_k' (t - t_k) \\ &= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{K-1} a_m h_k' p_\beta (t - mT - t_k) \end{aligned} \tag{1-10}$$

donde se construye el pulso $p_\beta(t)$ para que cumpla la condición $p_\beta(t)*p_\beta(-t)=p(t)$, siendo éste último un pulso de Nyquist. El receptor está adaptado al pulso $p_\beta(t)$ y por tanto, la señal detectada tiene la forma

$$\begin{aligned} r(t) &= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{K-1} a_m h_k' p_\beta (t - mT - t_k) * p_\beta (-t) = \\ &= \sum_{m=-\infty}^{\infty} \sum_{k=0}^{K-1} a_m h_k' p(t - mT - t_k) \end{aligned} \tag{1-11}$$

En el instante t=nT se desea detectar el símbolo representado por la amplitud a_n. La señal en ese instante será

$$r(nT) = r[n] = \sum_{m=-\infty}^{\infty} \sum_{k=0}^{K-1} a_m h_k' p[(n-m)T - t_k] \tag{1-12}$$

La componente de la señal de llegada al receptor a través del camino k=0 es exactamente $a_n h_0$, ya que se cumple la condición de Nyquist para los instantes (n-m)T con $t_0=0$, pero para los demás caminos, $p[(n-m)T-t_k]$ no es necesariamente nulo para cualquier cantidad n-m, y por lo tanto aparecerá, en general, interferencia intersimbólica. En el instante n la señal de llegada al

receptor será la suma de cada uno de los símbolos emitidos en los instantes anteriores, cada uno de ellos afectado de la amplitud

$$h_m = \sum_{k=0}^{K-1} h_k' p((n-m)T - t_k) \tag{1-13}$$

1.2.5. Modelo discreto de canal con interferencia intersimbólica

En general, habrá interferencia intersimbólica debida a todos los símbolos emitidos, pero en la práctica sólo se tomarán las N contribuciones de mayor amplitud, que se darán en los instantes m a m+N-1. Con esto se puede definir el canal discreto equivalente con propagación multicamino mediante la función de transferencia FIR siguiente:

$$h[n] = \sum_{k=0}^{N-1} h_k \delta[n-k] \tag{1-14}$$

y, sustituyendo la anterior expresión en la ecuación (1-12), la señal a la salida del filtro adaptado del receptor en el instante n tendrá la forma

$$r[n] = \sum_{k=0}^{N-1} h_k a_{n-k} = \sum_{k=0}^{N-1} h_k x[n-k] \tag{1-15}$$

Esta es la expresión general del canal lineal con propagación multicamino que se utiliza de aquí en adelante.

1.2.6. Distorsión no lineal y modelos de no linealidad

El problema de la distorsión no lineal en comunicaciones es una cuestión antigua y ha sido estudiada desde diferentes puntos de vista. Sus fuentes son varias, de las cuales la fuente más analizada es el transmisor.

Los amplificadores de potencia que se utilizan para comunicaciones en la banda de microondas son de dos tipos: unos son los amplificadores de tubo de onda progresiva (TWT), que presentan altos niveles de distorsión no lineal y que se utilizan cuando es necesario transmitir mucha potencia, como es el caso de las comunicaciones vía satélite. Cuando las potencias transmitidas son más bajas, se utilizan amplificadores basados en tecnología de efecto de campo (FET). Para aprovechar al máximo la potencia inyectada al transistor FET debe hacerse que éste trabaje en todo su margen dinámico; de esta manera, se consigue que el valor medio de la diferencia de potencial entre la fuente y el drenador sea lo menor posible y disipe la mínima potencia. El inconveniente de esto es la fuerte distorsión que se introduce en la señal de salida: el transistor FET es altamente no lineal [Minkoff, 1984]. Si la señal a transmitir se modula en PAM o en QAM, se producirá saturación en la constelación. Para evitarlo, se predistorsiona la señal de entrada al amplificador compensando la función de saturación de éste [González, 1997], [Karam, 1989], [Karam, 1991], [Pupolin/1, 1987], [Pupolin/2,

1987].

Sin embargo, el problema no puede solucionarse de la misma manera si es el canal el que introduce la saturación: en este caso, la compensación de canal desde el emisor no funciona bien, dado que no se puede conocer cual es el margen dinámico a la salida del canal y, además, éste es variante y dispersivo. La dispersión, unida a la saturación, produce efectos que no pueden ser eliminados fácilmente mediante predistorsión, tal y como veremos más adelante. A veces, las comunicaciones a alta velocidad necesitan de igualación no lineal para que sus prestaciones resulten aceptables [Mathews, 1991]: R. W. Lucky [Lucky, 1975] ya conjeturaba que en transmisiones de datos por el canal telefónico a velocidades mayores de 4800 baudios la tasa de fallos depende casi exclusivamente del comportamiento no lineal del canal.

En el modelo de canal que veremos en este trabajo se supone que existen dos elementos no lineales [Pagés, 1995], [Pagés, 1997]. El primero de ellos es F(·) y corresponde, por ejemplo, a la saturación del dispositivo de salida del emisor. A la salida del canal existe otra no linealidad, que corresponde a la saturación del canal propiamente dicha. Para un conjunto **x** de valores de entrada correspondientes a los datos supondremos que:

$$\left.\begin{array}{l} G'(x) > 0 \\ F'(x) > 0 \end{array}\right\}, \quad \forall x \in \mathbf{x}$$
$$F'(0) = G'(0) = 1 \qquad (1\text{-}16)$$

es decir, ambas funciones son continuas y monótonas crecientes en el dominio de la variable de entrada, lo que no supone restricciones respecto de los casos prácticos. Para posteriores aproximaciones al problema, supondremos también que la derivada de estas dos funciones en el origen es unitaria, es decir, se trata de funciones de compresión.

Una función de saturación para modelar dispositivos que presenten saturación debe ser [Cann, 80]:

- lineal para pequeña señal;
- asintótica a un nivel límite determinado;
- computable con pocas operaciones para no consumir excesiva potencia de cálculo;
- capaz de modelar codos de diferentes formas.

Un modelo de no linealidad utilizado en la práctica para simular canales es la función tangente hiperbólica [Van Vlek, 66]:

$$G(x) = \frac{1}{\tanh\left(\frac{1}{\xi}\right)} \tanh\left(\frac{x}{\xi}\right) \qquad (1\text{-}17)$$

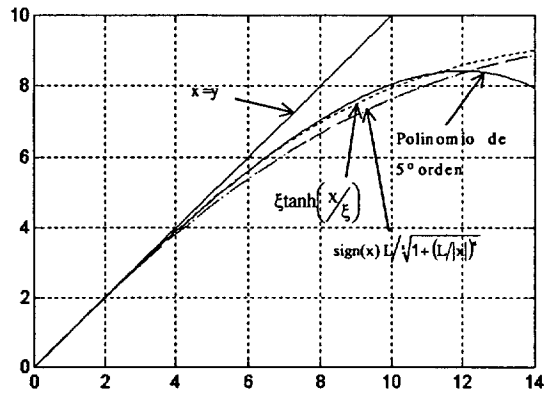


Figura 1-5. Modelos de no linealidad. s=3, L=ξ=10

ξ es el parámetro de saturación, que marca las cotas superior e inferior de la función, permitiendo que la derivada de la función no presente compresión en x=±1. La derivada en el origen no será unitaria en este caso, pero como $\{\tanh(1/\xi)\}^{-1}\cong\xi$, podemos hacer esta aproximación. Si se quiere que la derivada en el origen sea unitaria, el modelo puede ser:

$$G(x) = \xi\tanh\left(\frac{x}{\xi}\right) \tag{1-18}$$

Este modelo se aproxima bien a las funciones de saturación de los dispositivos de estado sólido, aunque no cumple la característica 4: el codo de la función no es modelable.

El modelo

$$G(x) = \operatorname{sign}(x)\, L\Big/\sqrt[s]{1+\left(L/|x|\right)^s} \tag{1-19}$$

es útil cuando se quiere caracterizar un dispositivo concreto ya que es ajustable por los parámetros s y L. Este modelo fue propuesto en [Cann, 80] y cumple las cuatro características.

En [Cann, 80] se hace una revisión de los modelos más utilizados. Uno de ellos consiste en la expansión de Taylor. Con este desarrollo se puede aproximar cualquier función, pero toda la información está concentrada en un punto y su entorno, y por ello para una buena aproximación, el orden ha de ser muy alto, con lo cual no cumple la característica 4. Además, no es "asintótica".

En la Figura 1-5 se puede ver una comparación de estos modelos, en los que suponemos que la saturación del canal se alcanza para amplitud 10 (L=ξ=10). El modelo basado en la tangente hiperbólica, que es el que se ha utilizado más en este trabajo está en línea discontinua de trazo largo. En trazo corto se observa el modelo de Cann. En trazo continuo hemos representado un polinomio de quinto orden de términos impares: aunque éste no es monótono

creciente, si lo será en todo el margen dinámico de la entrada y por ello constituye un modelo válido. Sin embargo, es muy difícil ajustarlo a un caso particular determinado.

1.3. MODELADO DE CANAL

La naturaleza no lineal en un canal de comunicaciones hace que éste no esté caracterizado por su respuesta al impulso. Esto significa que no es posible establecer un modelo general que incluya a todos los posibles sistemas no lineales, lo cual lleva a utilizar modelos no lineales que resultan mucho menos generales [Mathews, 1991].

Para hacer abordable el estudio analítico de sistemas en los que se presenta una no linealidad se construye un modelo en el que las componentes lineales del canal son separables de las no lineales. El modelo más general que llegamos a estudiar en este trabajo presenta la estructura de la Figura 1-6.

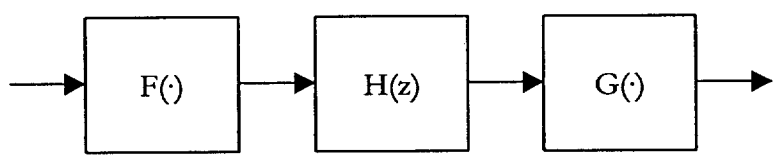


Figura 1-6. Modelo de canal no lineal empleado

1.3.1. Efecto de la parte lineal en la señal transmitida

H(z), con expresión en el tiempo h[n], es la "memoria" o parte lineal dispersiva del canal. El uso de un modelo invariante no representa restricciones en la mayoría de los análisis que se llevan a cabo en este trabajo. En general, la función de transferencia h[n,l] que se propone como modelo es causal y variante, de manera que:

$$h[n,l] = \sum_{k=0}^{N-1} h_k[l]\,\delta[n-k]. \qquad (1\text{-}20)$$

Para ver cuál sería el efecto del canal sobre la señal supóngase el canal lineal, invariante y de memoria finita N dado por la función

$$h[n] = \sum_{k=0}^{N-1} h_k \delta[n-k] \qquad (1\text{-}21)$$

y un receptor que toma un vector de entrada de M muestras consecutivas, donde será deseable que M≥N. La salida del canal en el instante n y en ausencia de ruido

puede escribirse en forma vectorial:

$$r[n]=\sum_{k=0}^{N-1}h_k x[n-k]=\mathbf{x}^T[n]\,\mathbf{h} \tag{1-22}$$

donde $\mathbf{h}=(h_0\cdots h_{N-1})^T$ corresponde al canal, y $\mathbf{x}[n]=(x[n]\cdots x[n-N+1])^T$ son las N muestras de la señal de la entrada al canal que presentarán contribución a la entrada al receptor en el instante n.

Si g[n] es el ruido aditivo presente, en un instante n determinado, el vector de M muestras $\mathbf{r}[n]=(r[n]\cdots r[n-M+1])^T$ a la entrada del receptor tiene la forma:

$$\mathbf{r}[n]=(r[n]\ \cdots\ r[n-M+1])^T+(g[n]\ \cdots\ g[n-M+1])^T=$$

$$=\begin{bmatrix}\sum_{k=0}^{N-1}h_k x[n-k]\\ \vdots\\ \sum_{k=0}^{N-1}h_k x[n-k-M+1]\end{bmatrix}+\begin{bmatrix}g[n]\\ \vdots\\ g[n-M+1]\end{bmatrix} \tag{1-23}$$

Se puede transformar esta ecuación a notación vectorial, quedando en la forma:

$$\mathbf{r}[n]=\begin{bmatrix}\mathbf{x}^T[n]\,\mathbf{h}\\ \vdots\\ \mathbf{x}^T[n-M+1]\,\mathbf{h}\end{bmatrix}+\begin{bmatrix}g[n]\\ \vdots\\ g[n-M+1]\end{bmatrix}=$$

$$=\begin{bmatrix}\mathbf{x}^T[n]\\ \vdots\\ \mathbf{x}^T[n-M+1]\end{bmatrix}\mathbf{h}+\begin{bmatrix}g[n]\\ \vdots\\ g[n-M+1]\end{bmatrix}= \tag{1-24}$$

$$=\mathbf{X}[n]\mathbf{h}+\mathbf{g}[n]$$

1.3.2. Centroides y clases en el espacio de datos de entrada al canal

Considérese como señal emitida una secuencia de un alfabeto en el que $x[n]$ puede tomar P valores equiespaciados y equiprobables: en ese caso el vector de dimensión N de entrada al receptor será $\mathbf{r}[n]$ y puede tener, en ausencia de ruido, P^{N+M-1} valores distintos, según se comprueba en la ecuación

(1-23): cada uno de los componentes de $\mathbf{r}[n]$ contiene N muestras; además, la secuencia de símbolos en los sumatorios de dos elementos consecutivos de la ecuación (1-23) diferirá en un valor. Como consecuencia de esto, el número de símbolos involucrado en el valor de $\mathbf{r}[n]$ será N+M-1. Como la señal de entrada tiene un alfabeto de P símbolos, el número total de combinaciones es P^{N+M-1}. A cada uno de estos valores posibles de $\mathbf{r}[n]$ se le puede denominar centroide[1].

Habrá P subconjuntos de centroides, cada uno de los cuales estará asociado a uno de los símbolos del alfabeto y contendrá $P^{N+M-1}/P = P^{N+M-2}$ centroides. A cada uno de estos conjuntos de centroides se le denomina clase. Si x_j representa al símbolo de cardinal j, la clase o grupo de centroides asociado al símbolo j se caracteriza porque $x[n]$ en la ecuación (1-23) es $x[n]=x_j$. Para el caso lineal, los P^{N+M-2} centroides de la clase j tienen la forma:

$$\mathbf{r}\left(x_i,\cdots,x^{(M+N-2)}\right)=\begin{bmatrix} x_j & \cdots & x[n-N+1] \\ \vdots & & \vdots \\ x[n-M+1] & \cdots & x[n-N-M+2] \end{bmatrix}\mathbf{h} \qquad (1\text{-}25)$$

donde $x[n]=x_j$, $x^{(k)} = x[n-k]$.

La tarea de decisión consiste en discernir a qué símbolo emitido en el instante n corresponde el vector de entrada recibido en ese instante, y para ello debe clasificarse cada vector como asociado a un determinado centroide o conjunto de centroides de la misma clase[2].

1.3.2.a Espacio de datos para el canal con N=2.

En la mayoría de las simulaciones se tomará el ejemplo sencillo de canal de fase mínima y con N=2. Para un canal de la forma $h[n]=\delta[n]+h_1\delta[n-1]$, es decir, N=2, y un receptor cuyo vector de muestras es de orden M=2, la entrada en ausencia de ruido tiene la siguiente expresión

$$\mathbf{r}[n]=\left(r[n]\;\; r[n-1]\right)^T=\begin{bmatrix} x[n] & x[n-1] \\ x[n-1] & x[n-2] \end{bmatrix}\begin{bmatrix} 1 \\ h_1 \end{bmatrix} \qquad (1\text{-}26)$$

Si la señal emitida es 8-PAM, por ejemplo, el número de centroides para todo el

---

[1] La señal recibida estará contaminada con ruido blanco gaussiano aditivo. Siempre que la potencia de éste sea mucho menor que la de la señal, los datos recibidos aparecerán agrupados en torno a cada uno de los valores que r[n] tendría en ausencia de ruido, razón por la cual a estos vectores se les llama centroides.

[2] Es decir, debe aislarse cada una de las clases, que estará formada por un determinado conjunto de centroides. Muchas veces, en aplicaciones de comunicaciones no resulta práctica la separación de cada uno de los centroides, ya que hay un gran número de ellos y además pueden no ser identificables debido al ruido. Es mucho más práctico identificar a qué símbolo pertenece cada agrupación de centroides, lo que es suficiente para clasificar la señal de llegada.

espacio de datos definido por $\mathbf{r}[n]$ es $P^{N+M-1}=8^{2+2-1}=512$ centroides. Para el caso en que $h_1=0.2$, la distribución en el espacio de observación del receptor presenta el aspecto de la Figura 1-7.

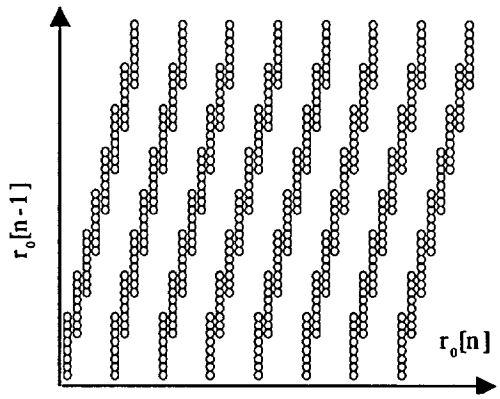


Figura 1-7. Conjunto de centroides para señal 8-PAM, canal $h[n]=\delta[n]+0.2\delta[n-1]$ (N=2) y receptor con vector de entrada de orden 2 (M=2).

En esta figura se observa que los centroides están agrupados en P=8 subconjuntos (los alineamientos casi verticales), cada uno de los cuales correspondiente a cada uno de los 8 símbolos del alfabeto.

1.3.3. Efecto de la parte no lineal en la señal transmitida

Véase ahora el efecto de introducir una no linealidad en el canal. Por sencillez se tratan por separado las dos linealidades comprendidas en el modelo de canal de la Figura 1-6. Se utiliza como ejemplo canal con N=2 del apartado 1.3.2.a.

1.3.3.a No linealidad sin memoria

Si el canal presenta una no linealidad a la entrada $F(\cdot)$ del tipo descrito en (1-16), la muestra presente en la entrada del receptor en el instante n será, en ausencia de ruido

$$\mathbf{r}[n]=\sum_{k=0}^{N-1} h_k F(x[n-k])=\begin{bmatrix} F(x[n]) \\ \vdots \\ F(x[n-N+1]) \end{bmatrix}^T \mathbf{h} \qquad (1\text{-}27)$$

En este caso la no linealidad afecta a cada uno de los datos por separado. Cuando ocurre esto se dice que el canal presenta una no linealidad sin memoria, y la compresión o disminución de distancias entre la clase en que está incluido el vector de datos $\mathbf{r}_0[n]$ sólo depende del símbolo emitido en el instante n.

En efecto, dos vectores de clases contiguas j=i y j=i+1 pero cuyos elementos x[n-k] son iguales para ambos, tienen la expresión siguiente:

$$\mathbf{r}\left(x_j,\cdots,x^{(M+N-2)}\right)=\begin{bmatrix} F(x_j) & \cdots & F(x[n-N+1]) \\ \vdots & \ddots & \vdots \\ F(x[n-M+1]) & \cdots & F(x[n-N-M+2]) \end{bmatrix}\mathbf{h} \quad (1\text{-}28)$$

La distancia euclídea entre los dos vectores será:

$$|h_0\{F(x_i)-F(x_{i+1})\}| \quad (1\text{-}29)$$

Como $\left.\frac{d}{dx}F(x)\right|_{x=x_i} > \left.\frac{d}{dx}F(x)\right|_{x=x_{i+1}}$ es inmediato ver que la distancia es menor para clases que tengan valores de x_i más altos. Esta distancia sólo depende de x_i y x_{i+1} o, lo que es lo mismo, del símbolo emitido en el instante n: Cuando la no linealidad está en la entrada del canal, la saturación afectará por igual a todos los elementos de una misma clase.

1.3.3.b No linealidad con memoria

Cuando el canal presenta una no linealidad a su salida G(·), la señal libre de ruido en el instante n será

$$r[n]=G\left(\sum_{k=0}^{N-1}h_k x[n-k]\right)=G\left(\begin{bmatrix} x[n] \\ \vdots \\ x[n-N+1] \end{bmatrix}^T \mathbf{h}\right) \quad (1\text{-}30)$$

Si se produce esta situación se dice que el canal tiene una no linealidad con memoria. En este caso, el valor de la compresión depende no solo del símbolo emitido en el instante presente n, sino que también depende de los N+M-2 símbolos emitidos, siendo N la longitud del vector **h** que caracteriza a la parte lineal del canal y M el número de muestras temporales que forman el vector de entrada al receptor.

Para comprobarlo se procede en la manera del apartado 1.3.3.a: se toman dos vectores de clases contiguas j=i y j=i+1 pero cuyos elementos x[n-k] son iguales para ambos. Su expresión tiene la forma

$$r\left(x_j,\cdots,x^{(M+N-2)}\right)=\begin{bmatrix} G\left(\begin{bmatrix} x_j \\ \vdots \\ x[n-N+1] \end{bmatrix}^T \mathbf{h}\right) \\ \vdots \\ G\left(\begin{bmatrix} x[n-M+1] \\ \vdots \\ x[n-M-N+2] \end{bmatrix}^T \mathbf{h}\right) \end{bmatrix} \tag{1-31}$$

La distancia euclídea para este caso será

$$\left| G\left(\begin{bmatrix} x_i \\ \vdots \\ x[n-N+1] \end{bmatrix}^T \mathbf{h}\right) - G\left(\begin{bmatrix} x_{i+1} \\ \vdots \\ x[n-N+1] \end{bmatrix}^T \mathbf{h}\right) \right| \tag{1-32}$$

La primera derivada de la función G(·) es decreciente; esta distancia será menor cuanto mayor sea el valor absoluto de las cantidades $\begin{bmatrix} x_j \\ \vdots \\ x[n-N+1] \end{bmatrix}^T \mathbf{h}$.

Cuando la no linealidad está a la salida del canal, la compresión afectará en mayor medida cuanto mayor sea el valor absoluto de la suma de los símbolos x[n-k].

1.3.3.c Efecto de la no linealidad sobre constelación 8-PAM

Las anteriores situaciones para 8-PAM, el canal visto y receptor con vector de entrada de dos muestras, están representadas en la Figura 1-8.

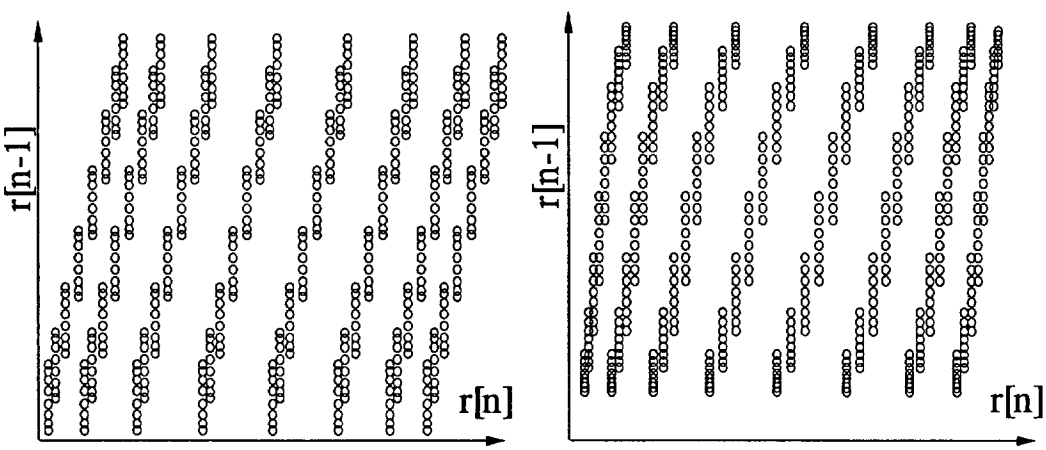


Figura 1-8. No linealidad a la entrada (izquierda) y a la salida.

Se tiene una señal PAM de 8 niveles equiprobables transmitida a través de un canal dispersivo de la forma $H(z)=1+0.2z^{-1}$. Se escoge como elemento no lineal del canal la función tangente hiperbólica

$$\frac{1}{\tanh\left(\frac{1}{\xi}\right)}\tanh\left(\frac{x}{\xi}\right)$$

descrita en la ecuación (1-17). El receptor toma vectores de dos muestras. Se ensaya su efecto sobre los datos colocando la no linealidad en la entrada y la salida del canal. En la Figura 1-8 se observa el conjunto de posibles vectores de orden dos de entrada al receptor. A la izquierda (no linealidad en la entrada) vemos que las agrupaciones correspondientes a las clases son paralelas, y que la distancia entre clases sólo depende de la dimensión r[n]. A la derecha (no linealidad en la salida) la distancia disminuye para los centroides que contienen símbolos para los que la cantidad $\left(x[n] \quad x[n-N+1]\right)$ **h** es mayor.

2. LA IGUALACIÓN EN COMUNICACIONES DIGITALES

2.1. FRONTERAS DE DECISIÓN

2.1.1. Frontera de máxima verosimilitud

La igualación puede interpretarse como el establecimiento de unas fronteras en el espacio de datos $\mathbf{r}[n]$ a la entrada del receptor que separan las diferentes clases. Sea la frontera i aquella que separa los datos de la clase i y los datos de la clase i+1: cada una de estas fronteras es una superficie formada por el conjunto de datos $\mathbf{r}_i$ que verifica una determinada ecuación

$$o_i(\mathbf{r}_i) = 0 \tag{2-1}$$

donde $o_i(\cdot)$ es cierta función escalar definida en el espacio de vectores de entrada.

Si se conoce esta ecuación, se puede establecer una regla de decisión de la siguiente manera: si x[n] es el símbolo emitido, se decide el valor de éste a partir del dato recibido $\mathbf{r}[n]$con el criterio de clasificar $\mathbf{r}[n]$ en la clase i (es decir, $x[n]=x_i$) si se verifica

$$\begin{aligned} o_j(\mathbf{r}[n]) &> 0; \qquad j \le i \\ o_j(\mathbf{r}[n]) &< 0; \qquad j > i \end{aligned} \tag{2-2}$$

La mejor frontera de decisión es no lineal, tal como se muestra mediante

el siguiente análisis clásico. Supóngase que se transmite la señal x[n] a través de un canal dispersivo, lineal, causal y de memoria finita igual a N representado por el vector $\mathbf{h} = (h_0 \cdots h_{N-1})^T$. Se dispone de un receptor que toma vectores de muestras de orden M en presencia de ruido aditivo n[n]. Se disponen P-1 fronteras de decisión $o_i(\mathbf{r}_i)=0$ (con $1<i<P-1$) de separación del conjunto de centroides correspondiente al símbolo i y al símbolo i+1 entre los P conjuntos de centroides.

Las fronteras o funciones de decisión serán óptimas para una determinada distribución de ruido n[n] si la probabilidad de error en la clasificación es mínima para esta distribución. Esta regla es la llamada de máxima verosimilitud[3]:

> La estimación $\hat{m}$ del símbolo emitido en el instante n es el símbolo que minimiza la probabilidad de error condicionada al dato **r**[n] o, lo que es lo mismo, el símbolo que maximiza la probabilidad de ocurrencia del símbolo emitido condicionada al dato recibido[4]. Según esto:

$$\hat{m} = \min_{x_i} P_e(x_i \mid \mathbf{r}) = \max_{x_i} P(x_i \mid \mathbf{r}) \tag{2-3}$$

Se indicará la densidad de probabilidad del dato **r**[n] por $f_r(\mathbf{r})$ y la probabilidad a priori de la ocurrencia del símbolo emitido por $P(x_i)$. Aplicando el Teorema de Bayes a la probabilidad a posteriori sobre la señal recibida:

$$\hat{m} = \max_{x_i} \frac{f_r(\mathbf{r} \mid x_i)P(x_i)}{f_r(\mathbf{r})} = \max_{x_i} f_r(\mathbf{r} \mid x_i)P(x_i) \tag{2-4}$$

La frontera de decisión $\mathbf{r}_i$ definida por la superficie $o_i(\mathbf{r}_i)=0$ debe situarse entre las clases i e i+1 de manera que

$$f_r(\mathbf{r}_i \mid x_i) P(x_i) = f_r(\mathbf{r}_i \mid x_{i+1}) P(x_{i+1}) \tag{2-5}$$

Aquí supondremos que las probabilidades a priori son iguales, con lo que la anterior ecuación queda en la forma:

$$f_r(\mathbf{r}_i \mid x_i) = f_r(\mathbf{r}_i \mid x_{i+1}) \tag{2-6}$$

2.1.2. Fronteras óptimas para ruido gaussiano blanco

En la ecuación (2-6) se ve la condición que debe cumplir la frontera de

---

[3] A. Fisher desarrolló en la década de 1920 el método de máxima verosimilitud para estimar los parámetros de una densidad de probabilidad. Entre otros, acuñó el término "función discriminante" para las fronteras de separación entre clases.

[4] Se supondrá siempre que la señal emitida tiene niveles equiprobables y equidistantes.

separación entre las clases i e i+1. Desarrollaremos esta ecuación para el caso en que el ruido a la entrada sea gaussiano. En la mayoría de los casos cada una de las muestras temporales tendrá una componente de ruido gaussiano blanco aditivo de potencia σ^2 que proviene de la misma fuente pero en instantes diferentes: por lo tanto, podemos suponer que cada muestra r[n] tiene una componente de ruido gaussiano independiente y de idéntica potencia que las otras. Excepto para x[n], cuyo valor es x_i o todos los símbolos x[n-k] para $1 \geq k \geq N+M-2$ tomarán cualquiera de los posibles valores del alfabeto.

Para cada símbolo $x[n]=x_i$, el vector de entrada al receptor (siempre que el canal sea lineal) **r**[n] toma cualquiera de los posibles valores de:

$$\mathbf{r}\left(x_i,\cdots,x^{(M+N-2)}\right)=\begin{bmatrix} x_i & \cdots & x[n-N+1] \\ \vdots & & \vdots \\ x[n-M+1] & \cdots & x[n-N-M+2] \end{bmatrix}\mathbf{h}+\mathbf{g}[n]= \tag{2-7}$$

$$=\mathbf{X}_i[n]\mathbf{h}+\mathbf{g}[n]$$

Expresión que se deduce inmediatamente a partir de (1-23). Como se ha dicho en la sección 1.3.2, existen P^{M+N-2} vectores de la clase i.

Para cada uno de estos vectores, la densidad de probabilidad de la entrada $\mathbf{r}\,|\,x_i$ condicionada a que $x[m]=x_i$ y $x[m-k]=x^{(k)}$ tiene la expresión:

$$f_r\left\{\mathbf{r}\left(x_i,\cdots,x^{(M+N-2)}\right)\right\}=\left(\frac{1}{\sqrt{\frac{2\pi}{M}}\sigma}\right)^{N+M-1} e^{\frac{\|\mathbf{X}_i\mathbf{h}-\mathbf{r}\|^2}{2\sigma^2}} \tag{2-8}$$

La probabilidad condicionada por $x[n]=x_i$ es, en virtud de la ley de la probabilidad total, la suma de cada una de las probabilidades condicionadas por cada una de las combinaciones de valores posibles para $x^{(1)}$ hasta $x^{(M+N-2)}$. Suponiendo que $x^{(k)}$ puede tomar los valores $x^{(k_j)}$, con $1 \geq k_j \geq P$ obtenemos la siguiente expresión para la densidad de probabilidad condicionada por $x[n]=x_i$:

$$P(x_i\mid \mathbf{r})=\sum_{k_1=1}^{P}\cdots\sum_{k_{M+N-2}}^{P}\left\{f_r\left\{\mathbf{r}\left(x_i,x_{k_1}\cdots,x_{k_{M+N-2}}\right)\right\}\prod_{j=1}^{M+N-2}P\left(x_{k_j}\right)\right\} \tag{2-9}$$

Si se acepta que los símbolos x_j son equiprobables:

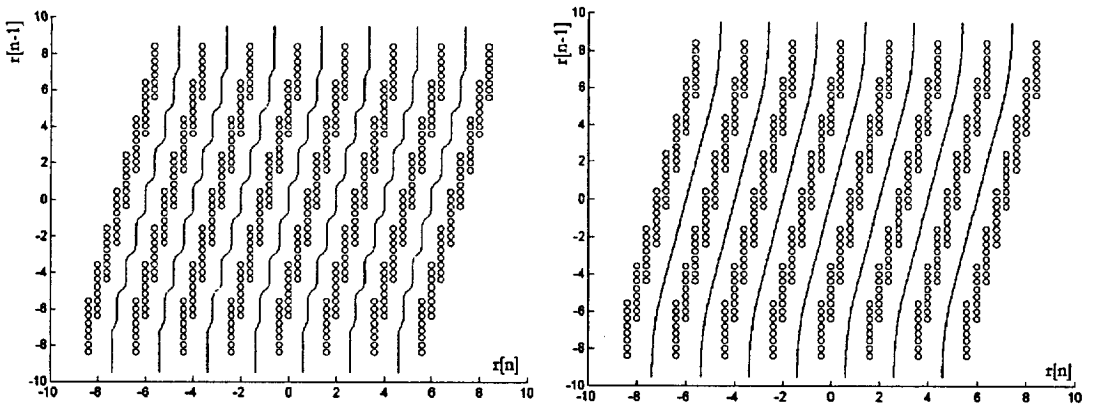


Figura 2-1. Fronteras MAP para SNR=30 (izquierda) y 10 dB.

$$P(x_i \mid r) = \frac{1}{P^{M+N-2}} \sum_{k_1=1}^{P} \cdots \sum_{k_{M+N-2}}^{P} f_r\left\{ r\left(x_i, x_{k_1} \cdots, x_{k_{M+N-2}}\right)\right. \tag{2-10}$$

La estimación de máxima verosimilitud es aquel valor de x_j que maximiza la anterior expresión (ver 2.1.1). Para obtener la frontera óptima, la condición que debe aplicarse es que en cada punto de esa frontera la probabilidad a posteriori de dos símbolos adyacentes sea igual. La decisión sobre el símbolo será (equivalentemente a la de la ecuación (2-3)), de la siguiente manera:

$$\text{decisión}: \quad d = \begin{cases} x_{j+1} & \text{si} \quad P(x_j \mid r) - P(x_{j+1} \mid r) > 0 \\ x_j & \text{si} \quad P(x_j \mid r) - P(x_{j+1} \mid r) < 0 \end{cases} \tag{2-11}$$

2.1.3. Fronteras de decisión para constelación 8-PAM, canal lineal.

La frontera de separación entre dos clases debe verificar que en todos sus puntos las densidades de probabilidad de las dos clases sea igual. Se puede ver en un ejemplo sencillo que esta frontera es altamente no lineal.

Se toma un canal cuya función de transferencia $H(z)=1+0.2z^{-1}$ y un receptor que toma datos de dimensión 2. Se calcula numéricamemte la frontera de máxima verosimilitud mediante el método de Bolzano[5].

---

[5] Para cada par de clases contiguas se escoge un intervalo en la dimensión r[n] que incluya los extremos de las dos clases en esta dimensión. Para cada valor del eje r[n-1] desde la parte inferior hasta la superior y a intervalos de 0.01 y para los extremos del intervalo r[n] se calculan las densidades de probabilidad de las dos clases según la ecuación (2-9). Se toma luego en el eje r[n] el valor que dé menor diferencia de densidades de probabilidad y el valor medio de los anteriormente utilizados. Se itera el proceso hasta que el intervalo en el eje r[n] sea menor que 0.01. De esta manera, se ha forzado una separación de puntos de 0.01 en el eje vertical y una precisión de 0.01 para cada intervalo en el eje horizontal.

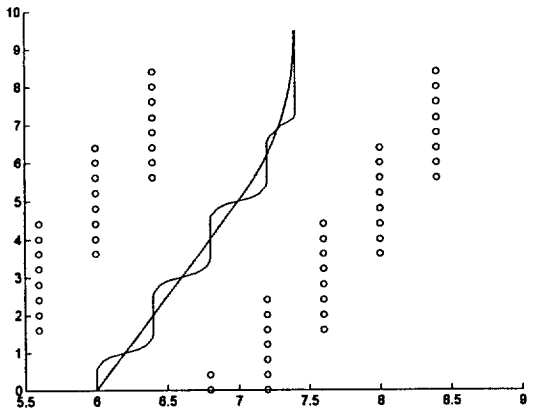


Figura 2-2. Detalle de las fronteras MAP para SNR=10 y 30 dB de la Figura 2-1

En la Figura 2-1 vemos que la frontera es no lineal y depende de la potencia de ruido. Adviértase también que para potencias de ruido altas, la frontera se acerca más a una frontera lineal debido a que las colas de las gaussianas que están más lejos toman más importancia en la suma y suavizan la intersección. Ello significa que cuando el ruido es alto, no es demasiado importante utilizar fronteras no lineales dado que el coste computacional quizá no justifique la mejora. Por otro lado, una buena aproximación es utilizar fronteras que presenten no linealidad en los extremos. Sin embargo, estas no linealidades no deben ser idénticas dado que para ciertos tipos de canales no lineales (según veremos más adelante en este capítulo) los extremos de la frontera óptima presentarán formas diferentes.

La Figura 2-2 muestra una porción de ambas fronteras de decisión para su comparación.

2.1.4. Separabilidad de las clases

La distribución de los datos en el espacio de entrada al receptor depende fuertemente de las no linealidades en el canal. Supóngase que se tiene un canal dispersivo determinado por la función de transferencia

$$H(z) = 1 + \sum_{k=1}^{N-1} h_k z^{-k}$$

es decir,

$$h[n] = x[n] + \sum_{k=1}^{N-1} h_k x[n-k]$$

Supónganse, además, el modelo de canal de la Figura 1-6 con las restricciones de la ecuación (1-16), (págs. 13 y 10). Parece obvio que, aunque en el caso lineal (es decir, con F(x)=G(x)=x) el conjunto es linealmente separable, la clasificación para el canal no lineal no tiene porqué ser realizable linealmente.

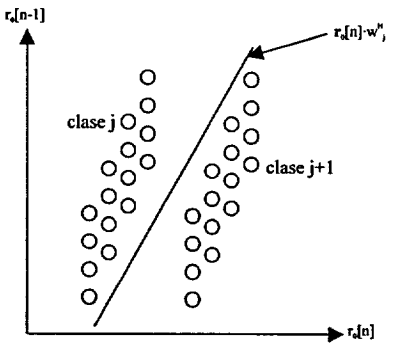


Figura 2-3. Separabilidad lineal de clases para el caso bidimensional con r[n]=(r[n] r[n-1]).

2.1.4.a Definición de separabilidad lineal de vectores:

El conjunto de posibles vectores

$$\{\mathbf{r}[n]\} = \{\mathbf{X}[n]\mathbf{h}^T\}$$

de dimensión M es linealmente separable en las clases x_i si para cada vector $\mathbf{r}[n]$ del espacio de datos a la entrada del canal perteneciente a la clase x_j y para algún vector $\mathbf{w}_i = [w_0 \ \cdots \ w_{M-1} \ \ w_M]^H$, de dimensión M+1 y definido en el espacio de datos, se cumple que:

$$\begin{bmatrix} \mathbf{r}[n] \\ 1 \end{bmatrix} w_i^H > 0 \quad j > i$$
$$\begin{bmatrix} \mathbf{r}[n] \\ 1 \end{bmatrix} w_i^H + w_M < 0 \quad j \le i \qquad (2\text{-}12)$$

La ecuación $\begin{bmatrix} \mathbf{r}[n] \\ 1 \end{bmatrix} \mathbf{w}_i = 0$ es el llamado hiperplano de discriminación[6] entre las clases x_i y x_j.

Se introducen ahora las componentes no lineales del modelo. Aún suponiendo que se cumpla la ecuación (1-16) para el conjunto $\mathbf{r}\{x_1, \ldots, x_n\}$, no se puede afirmar que se cumpla la separabilidad lineal de los elementos $\mathbf{r}$

$$\mathbf{r}_j \mathbf{w}_i^H > 0 \quad i \ge j$$
$$\mathbf{r}_j \mathbf{w}_i^H < 0 \quad i < j \qquad (2\text{-}13)$$

siendo

[6] Ver nota 3

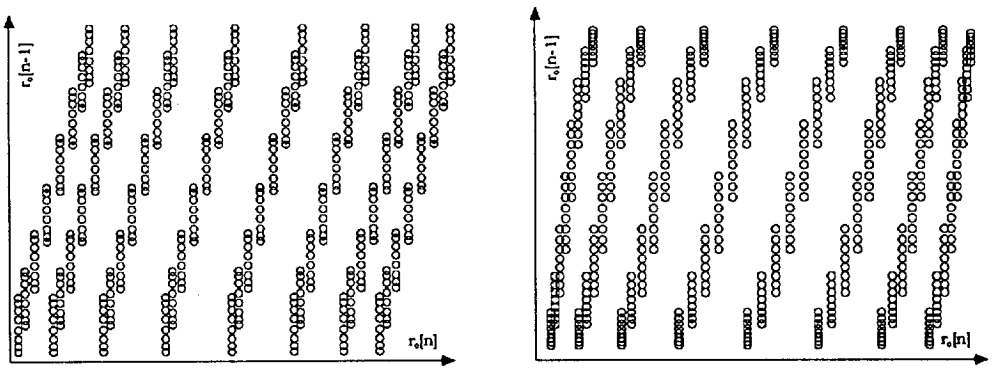


Figura 2-4. No linealidad a la entrada (izq.) y a la salida.

$$\mathbf{r}_j[n] = \begin{bmatrix} G\left(\left[F(m_j) \quad F\left(x[n-N+1] \right) \right] \mathbf{h} \right) \\ \vdots \\ G\left(\left[F\left(x[n-M+1] \right) \quad F\left(x[n-N-M+2] \right) \right] \mathbf{h} \right) \end{bmatrix} \qquad (2\text{-}14)$$

En una constelación PAM de P niveles con canal de orden dos, sin embargo, cuando las clases no se solapan, siempre son separables linealmente (véase Anexo 2.1).

2.1.4.b Espacio de datos para constelación 8-PAM

En la Figura 2-4 se observa el conjunto de posibles vectores de orden dos de entrada al receptor, para señal 8-PAM y canal $H(z)=1+0.2 \cdot z^{-1}$. En el primer caso, una buena aproximación es utilizar discriminantes lineales y paralelos, y si las clases no son separables linealmente, tampoco lo serán mediante cualquier otro método. En el segundo caso, los discriminantes no deben ser paralelos y pueden obtenerse mejores resultados si no son lineales. Las gráficas son para $L=\xi=7$.

2.2. COMPENSACIÓN DE CANAL

La compensación de canal consiste en acondicionar la señal desde el emisor para que su paso a través del canal produzca el mínimo efecto posible. Se trata, básicamente, de introducir una no linealidad que se aproxime a la inversa de la no linealidad del canal. En la Figura 2-5 se observa un ejemplo de no linealidad introducida por un amplificador de guía de onda progresiva. Existen tres vertientes de esta técnica: compensación estática, compensación adaptativa local y compensación adaptativa remota.

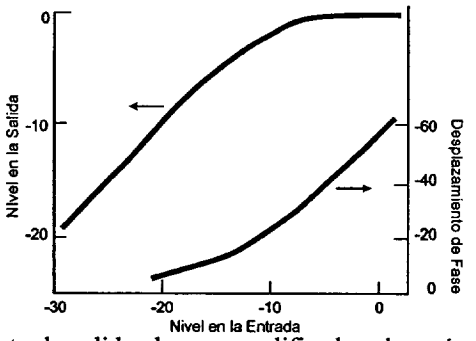


Figura 2-5. Característica entrada-salida de un amplificador de guía de onda progresiva (TWT) [Namiki, 1983]

2.2.1. Compensación estática

La compensación estática consiste en compensar la señal mediante una función fija que se aproxime a la inversa de la del elemento que introduce la distorsión. Ello presupone que la distorsión no varía con el tiempo. Un ejemplo sencillo se relata en [Nojima, 1980]. La hipótesis de invarianza es, en sentido estricto, incorrecta, ya que por lo menos deberá contarse con las derivas por envejecimiento, calor y variación de la tensión de alimentación.

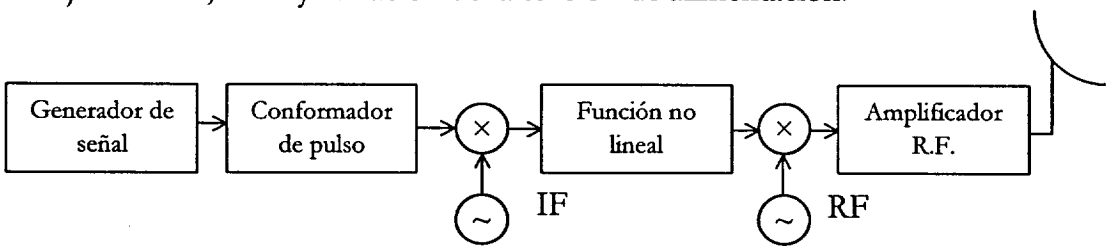


Figura 2-6. Compensación estática. La salida del conversor a IF pasa por una función no lineal que expande la señal a fin de compensar la compresión que sufrirá la señal en posteriores etapas de la transmisión.

La Figura 2-6 muestra el diagrama de bloques de la compensación estática.

2.2.2. Compensación adaptativa local

La componente no lineal del canal será, en general, no estacionaria, es decir, variará con el tiempo. La manera de seguir las variaciones de la forma de la no linealidad es compensar la distorsión de forma dinámica desde el emisor. Un ejemplo es [Lazzarin, 1994]: se supone que la distorsión está en el emisor y se toman muestras de la salida de éste para poder retroalimentar un sistema adaptativo de predistorsión de la señal.

El control de la compensación se hizo al principio mediante métodos analógicos y se ponía entre el modulador de radiofrecuencia y el amplificador. El tipo de compensación que se lleva a cabo íntegramente en el emisor se denomina predistorsión. Para controlar la compensación se utiliza el error entre la señal de

entrada al emisor y una muestra de la salida de la antena. Un esquema de aprendizaje del sistema es el algoritmo LMS en versión analógica, o método de la correlación.

La distorsión se puede modelar con un polinomio de tercer orden de la forma $f(y)=y+\xi|y|^2y$, siendo y la salida del modulador. Para compensar esta distorsión se puede utilizar una función también de tercer orden $g(x)=x+\eta|x|^2x$, que se ajusta de manera que la función $f(g(x)) \approx x$. Si se desarrolla esta expresión se obtiene

$$\begin{aligned} f(g(x)) &= g(x) + \xi|g(x)|^2 g(x) = \\ &= x + \eta|x|^2 x + \xi\left[\left|x + \eta x|x|^2\right|^2 \left(x + \eta x|x|^2\right)\right] \approx \\ &\approx x + (\eta + \xi)|x|^2 x \end{aligned} \tag{2-15}$$

La aproximación de la ecuación (2-15) es válida si $|x| \gg \eta|x|^2x$. El error de la compensación será

$$e = f(g(x)) - x = (\eta + \xi)|x|^2 x \tag{2-16}$$

y el gradiente de éste respecto del parámetro de ajuste η tiene la expresión

$$\frac{\partial e}{\partial \eta} = -\frac{\partial e \cdot e^*}{\partial \mathrm{Re}(\eta)} + j\frac{\partial e \cdot e^*}{\partial \mathrm{Im}(\eta)} = 2(\eta + \xi)|x|^6 \tag{2-17}$$

Sustituyendo (2-16) en (2-21) se obtiene la expresión simplificada

$$\frac{\partial e}{\partial \eta} = 2|x|^2 e \cdot x^* \tag{2-18}$$

Para hacer que el error tienda a su valor mínimo, se procede a variar η con el tiempo según la regla de actualización

$$\frac{\partial \eta}{\partial t} = -\lambda\frac{\partial e}{\partial \eta} = -\lambda|x|^2 e \cdot x^* \tag{2-19}$$

Es decir, se fuerza a que η varíe en la dirección contraria del gradiente del error.

En la práctica, se utiliza $E\{|x|^2\}$ en lugar de $|x|^2$. Al tener la esperanza valor constante, simplifica el algoritmo. Si se substituye $\lambda\, E\{|x|^2\}$ por γ en (2-19) y se integra ésta se obtiene:

$$\eta = -\gamma \int_{T_0}^{t} e \cdot x^* \, dt \tag{2-20}$$

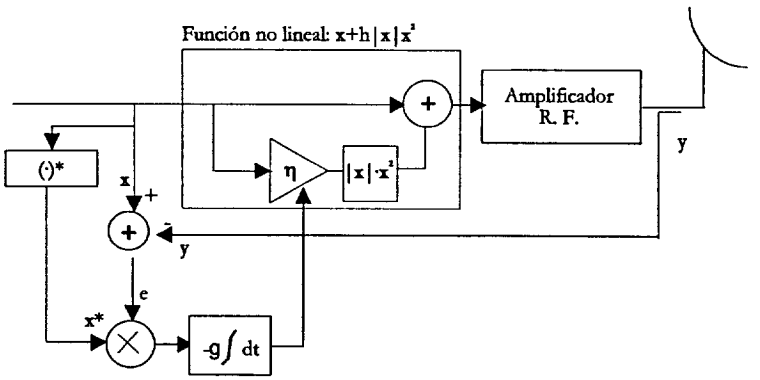


Figura 2-7. Compensación adaptativa local o predistorsión

La Figura 2-7 muestra la realización práctica de este algoritmo. La señal x(t) se introduce en el dispositivo no lineal de tercer orden. La salida pasa por el amplificador, obteniendo la señal distorsionada y(t). Con ella y con la señal de entrada se construye la señal error. Esta señal se multiplica por el conjugado de la entrada y el producto se integra. La salida del integrador controla la ganancia de un amplificador controlado por tensión, que finalmente amplifica la componente cúbica del dispositivo de compensación.

2.2.3. Compensación adaptativa remota

Una tercera y más elaborada solución consiste en suponer que es el canal el que distorsiona la señal, y en adaptar de forma dinámica el elemento de compensación de la distorsión introducida por aquél mediante retroacción desde el emisor. Esta aproximación tiene un inconveniente: existe un retardo entre la emisión de la señal y la retroacción debido al camino de ida y vuelta entre emisor y receptor: en el caso de que este retardo sea grande, afectará a las características del sistema respecto de la no estacionaridad del canal. Además, el filtro de predetección afectará al algoritmo ya que introduce distorsión lineal. El esquema de compensación remota analógica se muestra en la Figura 2-8.

Un esquema de compensación remota lleva a cabo el mismo algoritmo que un esquema de compensación adaptativa local, excepto porque el cálculo de η se hace en el receptor. Para calcular el error se utiliza la señal detectada: el sistema es dirigido por datos. Obsérvese que si el parámetro η varía lentamente con respecto a la velocidad de información a través del canal, entonces la potencia necesaria para reenviar este parámetro al emisor es pequeña: la antena utilizada puede ser de las típicas utilizadas en telemetría. Por supuesto, debe suponerse que el canal es estacionario, o que varía lentamente en relación con la capacidad del canal de retorno al emisor.

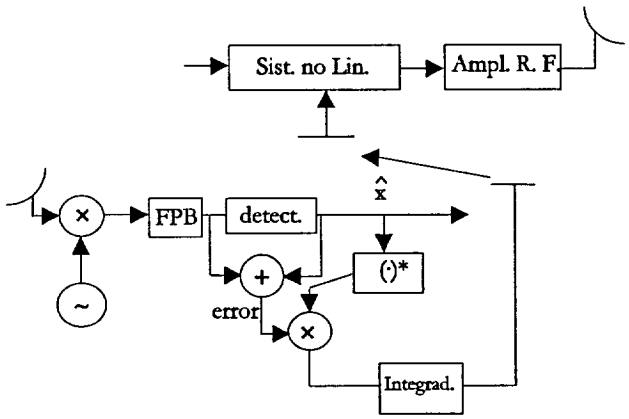


Figura 2-8. Compensación adaptativa remota

2.3. IGUALACIÓN LINEAL Y SUS LIMITACIONES

La igualación lineal consiste en una transformación lineal del vector de muestras de entrada de orden n hacia un espacio de orden 1 con el objeto de conseguir, a la salida, vectores de orden uno separables directamente en clases. Por ello, el clasificador lineal está limitado a problemas que sean (estrictamente) linealmente separables. Sin embargo, todos los problemas de clasificación con los que vamos a ensayar nuestros sistemas serán linealmente separables siempre que el canal sea de fase mínima y el ruido sea suficientemente bajo.

Otro de los inconvenientes del igualador lineal es que la transformación lineal no acondicionará adecuadamente las señales para que puedan ser separadas en clases: el plano de discriminación que forman el igualador lineal más los umbrales de detección no es el discriminador óptimo y, por tanto, una solución no lineal producirá un resultado más satisfactorio, como veremos en el siguiente apartado.

Una opción en igualación es la de ignorar la no linealidad del canal e igualar haciendo la aproximación de que el canal es lineal. En este caso se aplicará un algoritmo de filtrado lineal seguido de un dispositivo de decisión dura con intervalos homogéneos. Para ello basta con aplicar un filtro FIR adaptativo según la Figura 2-9, lo que constituye un control automático de ganancia, ya que los pesos del filtro tenderán a normalizar la señal para que se asemeje lo más posible a la señal deseada. El decisor multinivel siguiente debe tener simplemente sus umbrales equidistantes de las señales deseadas.

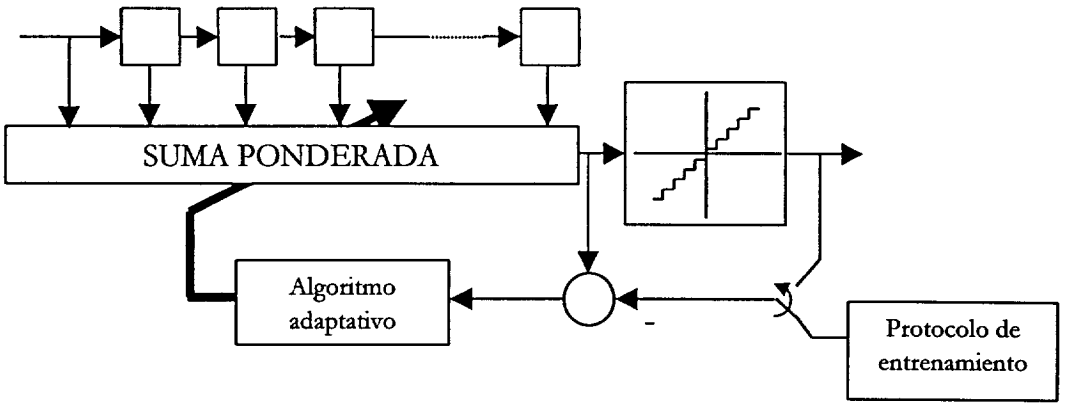


Figura 2-9. Igualación lineal multinivel

Se simula este sistema para el canal $H[z]=1+0.2z^{-1}$ y una no linealidad tipo tangente hiperbólica situada a la salida del canal. El filtro es de dos muestras y la señal de entrada PAM de 8 niveles. La Figura 2-10 muestra la igualación lineal en los casos en que hay no linealidad en la entrada del canal o la salida del canal. Las líneas en la figura representan los umbrales de decisión. Se observa que, a causa de la no linealidad del canal, el igualador no clasifica bien todas las señales o las sitúa peligrosamente cerca de los umbrales.

El sistema, como es sabido [Widrow, 1995], se aproxima a la solución óptima de Wiener, pero en la decisión se supone que las clases están equiespaciadas. Equivalentemente, se hace una proyección del espacio de salida en la dimensión horizontal y se discrimina mediante umbrales equiespaciados según la señal PAM del emisor. Dado que las señales no están equiespaciadas, la tasa de error en detección aumenta gravemente.

Este sistema no funcionará bien incluso para canales casi lineales, dado que un pequeño error en la disposición de los umbrales debido a la distorsión no lineal puede provocar un aumento considerable de la tasa de fallos. Obsérvese que las mayores fuentes de error se situarán en los bordes del conjunto de datos

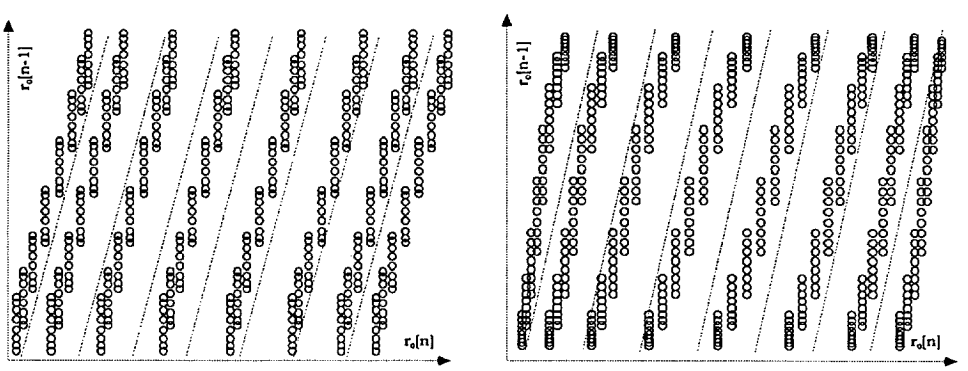


Figura 2-10. Espacio de señal de entrada al igualador cuando el canal presenta no linealidad a la entrada (izquierda) y a la salida del canal

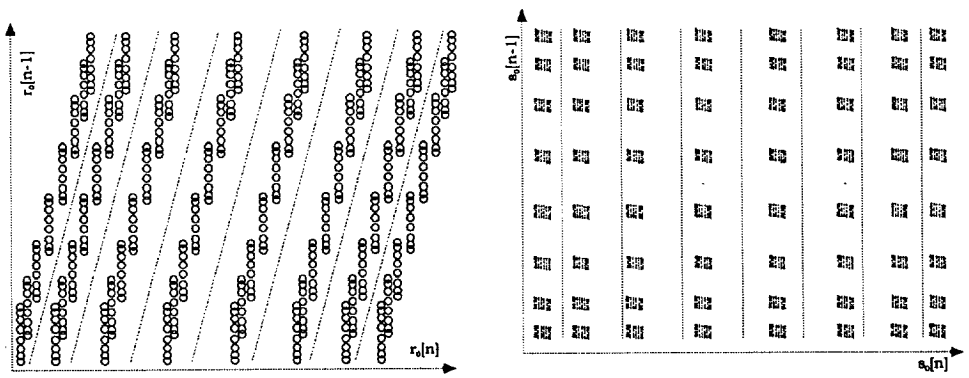


Figura 2-11. Igualación con umbrales adaptativos y no linealidad colocada a la entrada del canal. A la derecha se observa la salida del igualador en el instante n con respecto al instante n-1 para apreciar la baja correlación entre la dimensión $s_0[n]$ y la dimensión $s_0[n-1]$. Las líneas verticales en la derecha representan los umbrales de decisión sobre la dimensión $s_0[n]$.

de entrada.

Para solucionar el problema, una aproximación más fina que la anterior consiste en hacer que los umbrales de decisión sean adaptativos. En el caso en que la no linealidad esté en la entrada del canal, la aproximación del filtro FIR con umbrales adaptativos es una solución adecuada a la distribución de los datos (Figura 2-11). No obstante, como se puede ver en la Figura 2-12, esta aproximación sigue siendo mala ya que la distribución de los datos no sigue una estructura paralela , lo que produce errores en los bordes del conjunto de datos, donde la distorsión es mayor (véase la sección 1.3.3).

Una tercera aproximación es hacer los umbrales independientes unos de otros, como veremos más adelante, o utilizar fronteras no lineales, como se ve en el siguiente apartado.

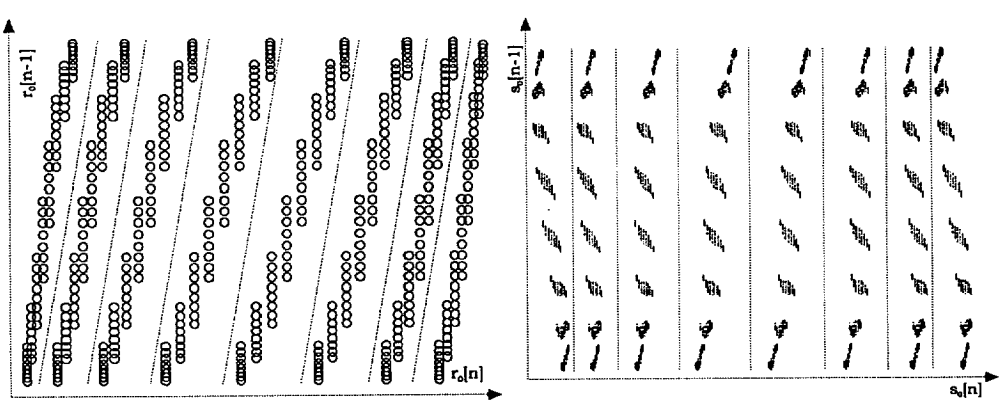


Figura 2-12. Igualación con umbrales adaptativos para el ejempo de la Figura 2-11, con una no linealidad con memoria. En los extremos, donde los datos están más cerca debido a la distorsión, la clasificación de éstos se ve comprometida. A la derecha puede observarse la salida del filtro en el instante n con respecto a la salida en el instante n-1. Existe una correlación más alta entre los datos que en el canal con no linealidad sin memoria.

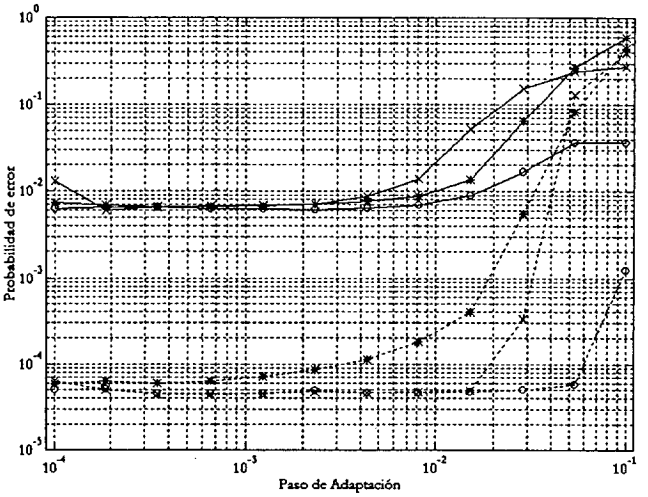


Figura 2-13. Simulaciones del filtro FIR con umbrales adaptativos para diferentes pasos de adaptación y para las funciones de coste cuadrática con activación lineal (*), cuadrática con activación tangente hiperbólica () y entrópica (o) (véase capítulo 3). Las gráficas en línea continua representan ensayos para SNR=25 dB y las gráficas en trazo discontinuo son para SNR=30 dB. No linealidad en la entrada.

2.3.1. Filtro FIR en canal con no linealidad sin memoria

Se simula una transmisión 8-PAM cuyo alfabeto es {±1 ±3, ±5, ±7}a través de un canal cuya respuesta al impulso es h[n] = δ[n]+0.2δ[n-1]. Se inserta una no linealidad en la entrada del canal con la forma

$$G(x) = \frac{1}{\tanh\left(\frac{1}{\xi}\right)} \tanh\left(\frac{x}{\xi}\right)$$

Se escoge ξ=10 y SNR= 25 y 30 dB y se simula el filtro FIR entrenado mediante LMS utilizando las funciones de coste cuadrática con activación lineal, cuadrática con activación tangente hiperbólica y función de coste entrópica (véase Capítulo 3 para una presentación de dichas funciones de coste). Las comparaciones se llevan a cabo para diferentes pasos de adaptación μ. El intervalo utilizado para μ va desde 10^{-4} hasta 10^{-1} con valores espaciados en la escala logarítmica.

La Figura 2-13 corresponde a la tasa de errores de la simulación. Se observan buenas prestaciones, como es de esperar, dada la disposición paralela de las clases.

2.3.2. Filtro FIR en canal con no linealidad con memoria

Se ensaya el mismo sistema que en el apartado 2.3.1 pero colocando ahora la no linealidad en la salida del canal. Se puede observar claramente en la Figura 2-14 la degradación de las prestaciones en comparación con la anterior simulación.

2.4. IGUALACIÓN NO LINEAL

2.4.1. Introducción

El uso de igualadores no lineales ha quedado justificado en el apartado 0. Está claro que, aún cuando el canal sea lineal, la frontera de decisión óptima es no lineal, y más aún cuando el canal es de fase no mínima[7]. Sin embargo, cuando la probabilidad de error es baja, se opta por la igualación lineal para no penalizar la complejidad de cálculo en el procesamiento de la señal. Cuando la probabilidad de error obtenida por métodos lineales es muy baja, en igualación se puede optar por aumentar la dimensión del espacio de datos. Sin embargo, esta solución

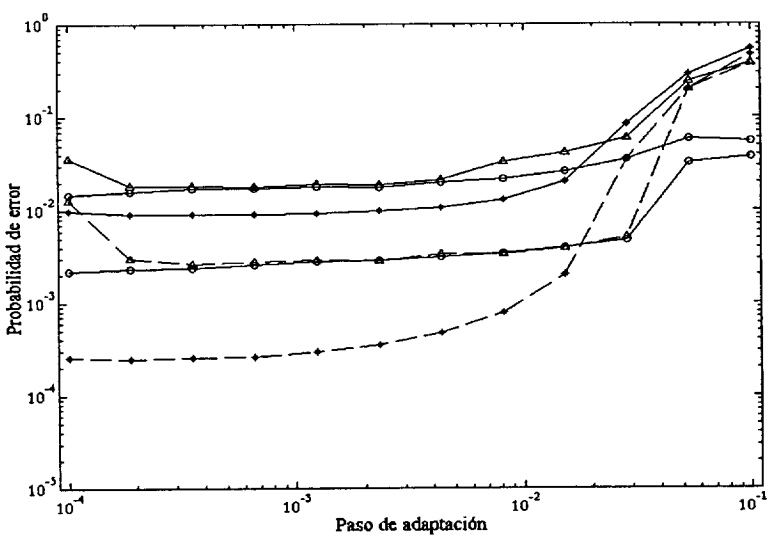


Figura 2-14. Simulaciones del filtro FIR con umbrales adaptativos para diferentes pasos de adaptación y para las funciones de coste cuadrática con activación lineal (*), cuadrática con activación tangente hiperbólica () y entrópica (o) (véase capítulo 3). Las gráficas en línea continua representan ensayos para SNR=25 dB y las gráficas en trazo discontinuo son para SNR=30 dB. No linealidad en la salida.

[7] Sin embargo, en esta tesis sólo se tienen en cuenta canales de fase mínima y señales multinivel, lo que ya

puede no ser suficiente. Entre otros problemas, si se emplea un algoritmo de gradiente para no aumentar excesivamente la complejidad (los algoritmos bloque tienen el inconveniente de necesitar una potencia de cálculo que crece exponencialmente con la dimensión del espacio de datos), se corre el peligro de disminuir la velocidad de convergencia. Esto es particularmente grave si se necesita una buena velocidad de seguimiento frente a canales no estacionarios. Por otro lado, los procesos basados en la solución de Wiener obtienen soluciones óptimas en presencia de ruido gaussiano, pero cuando el ruido no es gaussiano, la solución puede ser subóptima. Una solución basada en la incorporación de no linealidades puede proporcionar mejores prestaciones [Haykin, 1991].

La igualación no lineal es una solución que solventa los problemas de la igualación lineal pero que resulta mucho más compleja. Cuando se escoge igualar mediante métodos no lineales es necesario justificar esta solución en términos de la relación prestaciones/complejidad. Estas son las razones que, a primera vista, según nuestro entender, pueden hacer al diseñador desistir de utilizar procesado no lineal en igualación. Sin embargo, hay ya muchos trabajos publicados al respecto, de los cuales hacemos aquí una pequeña revisión.

Entre las soluciones no lineales a la igualación que existen se encuentran las siguientes [Weruaga, 1994]:

- igualadores que utilizan un filtro lineal combinado con una tabla de actualización ("look-up table", LUT)
- igualadores basados en estructuras no lineales clásicas como los filtros de Wiener y Hammerstein o los polinomios de Volterra;
- igualadores basados en redes neuronales.

2.4.2. Igualación mediante tablas de actualización (LUT).

Los sistemas de igualación basadas en tablas de actualización ("look-up tables") [Smith, 1988], [Yamazaki, 1991], [Weruaga, 1991], [Weruaga, 1994] se basan en una tabla de memoria. Si se tiene una transmisión de símbolos en un alfabeto de P niveles y se considera que el canal contiene N elementos significativos de memoria (esto es, el canal se modela por un filtro FIR de N coeficientes), el número de posibles amplitudes recibidas será de P^N, las cuales dependerán de los N últimos símbolos transmitidos. Si se toman vectores de muestras de dimensión M, se tendrán P^{N+M-1} vectores diferentes. Estos P^{N+M-1} elementos pueden almacenarse en una memoria de igual longitud. Si se conoce la combinación de símbolos emitida, se puede predecir cuál va a ser la señal de salida del canal. Esto hace que este método sea especialmente válido para cancelación de eco en transmisión de datos por bucle de abonado, ya que se conocen en todo momento

---

de por si justifica el uso de fronteras de decisión no lineales.

los componentes del eco. Si el eco es no lineal, simplemente es necesario almacenar en la memoria el valor de la respuesta no lineal a cada combinación de N+M-1 datos.

2.4.3. Igualación directa mediante polinomios de Volterra

El uso de series de Volterra ha sido muy popular [Billings, 1984], [Koh, 1985], [Sicuranza, 1985], [Sicuranza, 1986], [Biglieri, 1988], [Callender, 1992], [Carini/1, 1995], [Carini/2, 1995], [Fnaiech, 1995], [Im, 1995], [Nowak, 1996] porque ofrecen una cierta sencillez de manejo a la vez que buenas prestaciones, mientras que la dificultad de otros modelos más generales ha imposibilitado una exploración profunda de sus posibilidades. Se presenta aquí un resumen teórico acerca de la expansión en serie de Volterra extraído fundamentalmente de [Priestley, 1988], más un resumen de técnicas de igualación mediante el uso de series truncadas de Volterra, aunque una exposición más detallada puede encontrarse en las muy conocidas referencias [Schetzen, 1980] y [Mathews, 1991]. Por último se hace referencia a trabajos recientes en el tema.

Un sistema es lineal e invariante en el tiempo si su salida respecto de una entrada dada puede expresarse por la convolución de ésta con la respuesta h[n] del sistema al impulso unitario $\delta[n]$. Este sistema será causal si $h[n]=0$ para $n<0$. Por otro lado, la contribución de la entrada en el instante n_0-N, siendo n_o el instante actual, está determinado por el valor de h[n]. Por ello decimos que h[n] constituye la memoria del sistema.

Un sistema no lineal sin memoria puede ser representado por la curva y[n]=f(x[n]), y esta puede ser expandida por la serie de Taylor:

$$y[n] = \sum_{-\infty}^{\infty} c_m x^m [n] \qquad (2\text{-}21).$$

siendo c_m son los coeficientes de Taylor. En el caso en que el sistema tenga memoria, en principio, todo lo que podemos decir es que éste tiene una expresión en función de la entrada en cada uno de los instantes, de la forma:

$$y[n] = h(x[n], x[n-1], x[n-2], \ldots) \qquad (2\text{-}22)$$

Se supone siempre que el sistema es causal. Si esta función está suficientemente bien formada (es derivable infinitamente en el entorno de un punto), podrá ser expandida en serie de Taylor en el entorno de este punto. La expresión de esta serie en el origen $\mathbf{0}=(0, 0, 0, \ldots)$ es la siguiente:

$$y[n] = h_0 + \sum_{m_1=0}^{\infty} h_1[m_1]\, x[n-m_1] + \cdots$$
$$+ \sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty} h_2[m_1, m_2]\, x[n-m_1]\, x[n-m_2] + \cdots \qquad (2\text{-}23)$$
$$+ \sum_{m_1=0}^{\infty}\sum_{m_2=0}^{\infty}\sum_{m_3=0}^{\infty} h_3[m_1, m_2, m_3]\, x[n-m_1]\, x[n-m_2]\, x[n-m_3] + \cdots$$

y se denomina serie (discreta) de Volterra en honor a Vito Volterra, quien introdujo en 1870 esta expansión como generalización del desarrollo de Taylor de una función. El equivalente continuo se obtiene reemplazando los sumatorios de (2-24) por integrales.

Los elementos h_p son:

$$h_0 = h(0)$$
$$h_p[m_k, m_l, \ldots] = \left(\frac{\partial^p h}{\partial x[n-m_k]\, \partial x[n-m_l] \cdots} \right) \qquad (2\text{-}24)$$

y se denominan núcleos de Volterra[8].

Para realizar un filtro no lineal con una serie de Volterra es preciso truncarla. El esquema asociado a una serie de Volterra de segundo orden puede verse en la Figura 2-15.

El modelo de Volterra puede hacerse adaptativo, utilizando para ello los mismos algoritmos que para el caso lineal. Si se utiliza para la adaptación de los coeficientes el algoritmo LMS [Koh, 1985] y el objetivo a minimizar es el error cuadrático medio, ha de hallarse el gradiente de éste. Para los términos lineales, la expresión de la actualización de los pesos h_i es la habitual:

$$\Delta h_1[m_i,n] = \mu_1 e[n] x[n-m_i] \qquad (2\text{-}25)$$

---

[8] Norbert Wiener utilizó esta expansión como base para su estudio de sistemas no lineales, aunque trabajó en tiempo continuo, y su principal objetivo era encontrar transformaciones de esta serie cuyos términos fuesen ortogonales. Esta extensión fue presentada por Wiener en 1949 cuando trabajaba en el Instituto Nacional de Cardiología de Méjico. En una carta enviada a J. B. Wiesner, director del Research Laboratory of Electronics, MIT, decía: "*Le envío cierta información acerca de lo que pienso sobre circuitos no lineales y su prueba. El instrumento puede realizarse, y Lee puede diseñarlo. Teóricamente, no hay nada nuevo, pero creo que como técnica de ingeniería está al rojo vivo*". Acompañaba a la carta un informe titulado "Las Propiedades Características de los Sistemas Lineales y No lineales". Wiener continuó el estudio de su teoría y en 1958 publicó sus mayores contribuciones en [Wiener, 1958]. (Cfr, [Schetzen, 1981]). En esta última referencia puede encontrarse una introducción a esta teoría.

Para los demás términos la expresión se obtiene de manera similar. Por ejemplo, los términos cuadráticos tienen una actualización según la expresión:

$$\Delta h_2[m_i,m_j,n]=\mu_2 e[n]x[n-m_i]x[n-m_j] \qquad (2\text{-}26)$$

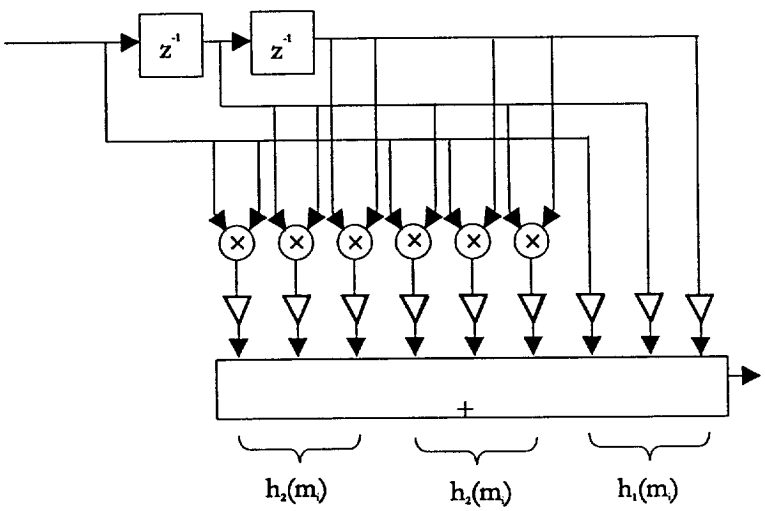


Figura 2-15. Filtro de Volterra de orden 2 para vectores de entrada de orden 3

También puede aplicarse el algoritmo RLS[9], consiguiendo unos resultados mucho mejores en términos de convergencia que con el LMS [Matthews, 1991], pero con un coste computacional mucho más elevado: si, utilizando el algoritmo LMS con N elementos de memoria para la entrada, se necesita un número de operaciones proporcional a N^2, utilizando el algoritmo RLS el número de operaciones es proporcional a N^4.

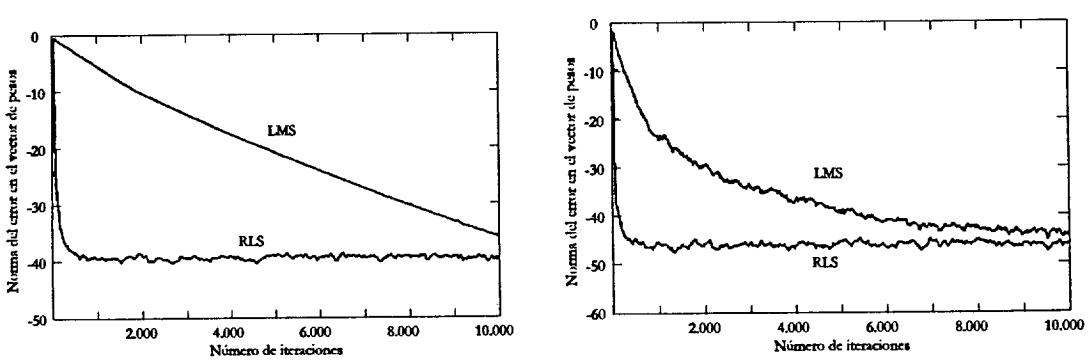


Figura 2-16. Comparación de la velocidad de convergencia del filtro de Volterra (ver texto) con algoritmo RLS y algoritmo LMS. Puede verse la comparación de la convergencia de los términos lineales (izquierda) y los términos cuadráticos.

[9] Para una exposición del algoritmo RLS, ver [Haykin, 1991].

La Figura 2-16 muestra la comparación de los dos algoritmos en cuanto a velocidad de convergencia. Los datos del emisor se filtran a través de un FIR de primer orden y fase mínima, y se procesan con un filtro de Volterra de segundo orden y entrada de dimensión 4 (es decir, consta de 4 núcleos de primer orden y dieciséis de segundo; sin embargo, por simetría, se considera que $h_2[m_i m_j] = h_2[m_j m_i]$ y por lo tanto los elementos de segundo orden se reducen a 10).

2.4.4. Igualación directa mediante redes neuronales

El nombre que se le ha dado a las redes neuronales viene de que su funcionamiento y (sobre todo) su estructura tienen cierto parecido con aquellos del tejido nervioso de los animales o, por lo menos, de lo que se cree saber acerca de éste[10]. Una red neuronal es un procesador distribuido, donde el procesado de los datos de las entradas se lleva a cabo en paralelo en cada una de las neuronas de que consta la red, que están distribuidas en varias capas. La neurona artificial es un dispositivo que consta de diversas entradas (dendritas) y una salida (axón), que es una función no lineal de las entradas. La red neuronal está formada por la interconexión de las salidas de cada una de las capas de neuronas con las entradas de la capa siguiente.

Algunas de las ventajas de las redes neuronales son:

- una neurona es, básicamente, un dispositivo no lineal, ya que la salida de ésta es una función no lineal de la suma ponderada de las entradas;
- proporcionan mapeado de entrada-salida. Esto significa, en particular, que, para un determinado entrenamiento adecuado, una red neuronal es capaz de clasificar determinadas entradas en clases diferentes;
- las redes neuronales tienen capacidad de aprendizaje. Por ello, además, las redes neuronales son capaces de adaptarse a los cambios del entorno en el que operan.

Existe una gran cantidad de estudios y ensayos que hacen referencia tanto a la estructura que debe tener la red dependiendo de la aplicación como a los algoritmos de entrenamiento de ésta, aunque no hay una teoría general que permita diseñar la red neuronal que más se ajusta a un determinado problema de clasificación. En cualquier caso, una descripción exhaustiva de técnicas de procesado neuronal está muy por encima del alcance y objetivos de esta tesis. Las referencias citadas [Zurada, 1992], [Haykin, 1994] son dos importantes elementos bibliográficos en los que puede encontrarse una descripción general de las redes neuronales.

Tres técnicas generales han sido aplicadas a la igualación mediante redes neuronales [Cid, 1994]. La primera de ellas es la basada en el perceptrón

[10] Sin embargo, las redes neuronales son sólo un muy rudimentario modelo del sistema nervioso.

multicapa (MLP) [Chen, 1990], [Mulgrew, 1991], [Bernardini, 1993], [Adali, 1997] . Otra se basa en redes de funciones de base radial (RBF) [Chen, 1991], [Mulgrew, 1991]. La última es la llamada red autoorganizativa, y fue propuesta por Teuvo Kohonen hacia 1981 (Véase Kohonen, 1990). Otros tipos de redes son las redes recurrentes (Elman, Jordan, Hopfield), que incluyen retroacción desde la capa de salida o una capa oculta.

2.4.4.a El perceptrón multicapa (MLP)

El elemento básico del MLP es el nodo o neurona sencilla. Se trata de un elemento que verifica la función

$$o = f(\mathbf{w}^H \mathbf{x}) \tag{2-27}$$

donde **w** es el vector de ponderaciones o pesos, y **x** es el vector de entrada a la neurona. f(.), llamada función de activación., suele ser una función de saturación, y nosotros nos restringiremos a una única función, la tangente hiperbólica o sigmoide bipolar:

$$f(x) = \frac{1 - e^{-x}}{1 + e^{-x}}. \tag{2-28}$$

En la Figura 2-17 puede verse gráficamente la estructura del nodo y la función de activación sigmoidal.

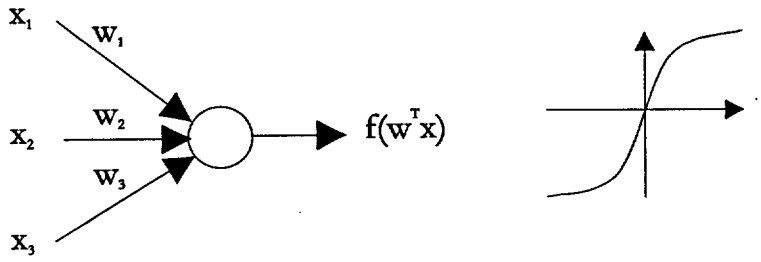


Figura 2-17. Neurona o nodo del perceptrón multicapa y función de activación sigmoidal.

En un MLP se disponen diversas capas de las neuronas descritas. La entrada se aplica a cada una de las neuronas de la primera capa, y la salida de cada una de éstas a la capa superior. El esquema correspondiente es el de la Figura 2-18. Entre los nodos de una misma capa no existen conexiones. La última capa proporciona las salidas necesarias para obtener la clasificación de los datos. Así, un clasificador de señales binarias constará de una única salida en la última capa y para señales N-PAM serán necesarias N/2 salidas.

El perceptrón multicapa se entrena según la regla de retropropagación.

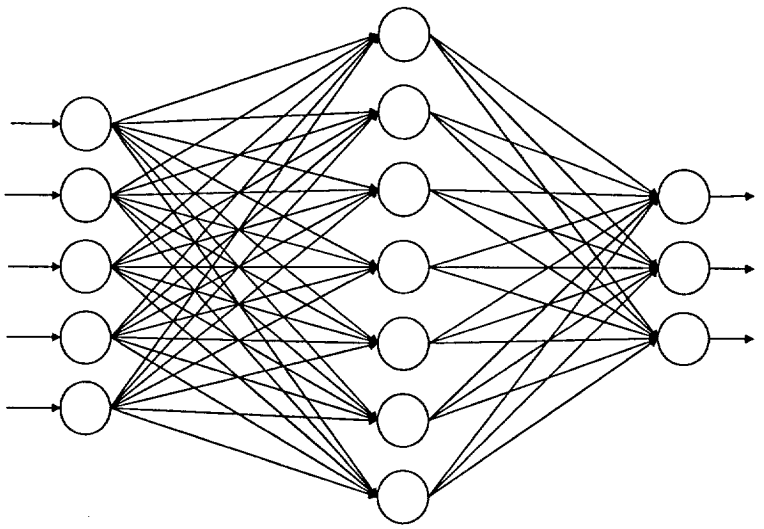


Figura 2-18. Estructura de un perceptrón multicapa en el que se observa una sola capa oculta con siete nodos y una capa de salida con tres nodos.

Esta regla fue introducida por P. J. Werbos[11] en 1975. Éste es un algoritmo de gradiente y se basa en la minimización de una función de coste de la salida o salidas de la red neuronal. Puede encontrarse una exposición general en [Widrow, 1990] o [Widrow, 1995].

Supóngase que se tiene una estructura neuronal con L capas. La salida de la neurona i de la capa l-1 es o_i^{l-1}, y el peso por el que se multiplica esta salida para introducirla en la neurona i de la capa l se denomina w_{ji}^{l} (véase la Figura 2-19) El vector de pesos de la red es **w**. Se define, asimismo, la salida deseada en la neurona j de la capa de salida como d_j. Si se utiliza como objetivo la minimización del error MSE, se define la función de coste como:

$$C(\mathbf{w}) = \sum_{j=1}^{N_L} \frac{1}{2}\left(d_j - o_j^L\right)^2 \qquad (2\text{-}29)$$

Minimizar esta función equivale a escoger los pesos adecuados para anular el gradiente del coste respecto de éstos, es decir, debe actualizarse cada uno de los

[11] El algoritmo de retropropagación fue publicado por Rumelhart en 1986 (...). De hecho, el entrenamiento por retropropagación fue descubierto independientemente en otros dos lugares casi al mismo tiempo (Parker, 1985; LeCun, 1985). Después del descubrimiento del algoritmo de retropropagación en la mitad de la década de los 80, se vio que este algoritmo había sido descubierto anteriormente por Werbos en su Tesis Doctoral en la Universidad de Harvard en agosto de 1974. La Tesis de Werbos fue la primera descripción documentada de cálculo de gradiente en modo inverso aplicada a modelos generales de redes, con las redes neuronales como un caso particular. Desgraciadamente, el trabajo de Werbos permaneció casi desconocido para la comunidad científica durante más de un decenio (Cfr. [Haykin, 1994]).

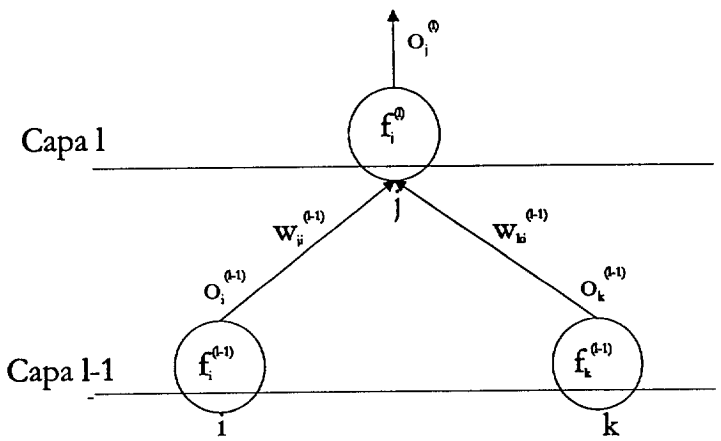


Figura 2-19. Notación utilizada en el perceptrón multicapa [Figueiras, 1998].

pesos de la red en sentido contrario al gradiente del coste, para que éste tienda a su mínimo. El vector que apunta a la dirección de máxima pendiente del coste respecto de sus pesos es:

$$-\frac{\partial C}{\partial w_{ji}^{l}} = -\frac{\partial C}{\partial o_{j}^{l}}\frac{\partial o_{j}^{l}}{\partial w_{ji}^{l}} = -\frac{\partial C}{\partial o_{j}^{l}} o_{i}^{l-1} f'^{l}_{j} \qquad (2\text{-}30)$$

siendo f(x) la función de activación (tangente hiperbólica) descrita anteriormente.

Esta expresión depende de la derivada del coste respecto de las salidas de la capa inmediatamente anterior. Si se desarrolla esta derivada se obtiene la siguiente solución:

$$\frac{\partial C}{\partial o_{j}^{l}} = \sum_{n_{l+1}=1}^{N_{l+1}} \frac{\partial C}{\partial o_{n_{l+1}}^{l+1}} \frac{\partial o_{n_{l+1}}^{(l+1)}}{\partial o_{j}^{l}} = \sum_{n_{l+1}=1}^{N_{l+1}} \frac{\partial C}{\partial o_{n_{l+1}}^{l+1}} w_{n_{l+1}}^{l+1} f'^{l+1}_{j} \qquad 1 \le l < L \qquad (2\text{-}31)$$

Según este algoritmo, debe empezarse por calcular los gradientes del coste respecto de las neuronas de salida, que serán $\frac{\partial C}{\partial o_{j}^{L}} = -(d_j - o_j^L)$. Con estos gradientes, se computa el coste respecto de las salidas de la capa anterior mediante la ecuación (2-31) y se itera el proceso con los resultados obtenidos para calcular los gradientes respecto de las salidas de la capa anterior, hasta llegar a la capa de entrada. Después se procede al cálculo del gradiente del coste respecto de los pesos mediante la ecuación (2-30), para finalmente actualizarlos.

El algoritmo de retropropagación es una generalización del algoritmo LMS, que usa un gradiente instantáneo con el método de máxima pendiente para minimizar el error cuadrático medio. Sin embargo, el error cuadrático no es una función cuadrática de los pesos. La naturaleza desconocida de este error cuadrático medio dificulta la predicción de la velocidad de convergencia del

algoritmo. Además, existen mínimos locales. Esto no pasa con los filtros de Volterra, cuyo comportamiento es más simple y, en muchos aspectos, más predecible [Widrow, 1995]. La dificultad para garantizar que se alcance el mínimo global, el tiempo necesario para la convergencia constituyen importantes inconvenientes en el caso de igualación de canal. Por otra parte, la potencia de cálculo del dispositivo que lleva a cabo el proceso MLP debe ser alta, dado que el coste computacional del algoritmo es elevado.

2.4.4.b Redes de funciones de base radial

La red descrita en el apartado 2.4.4.a se basa en el cálculo de los productos escalares del vector de entradas por el vector de pesos de cada neurona. Otro tipo de redes, basadas en las llamadas Funciones de Base Radial (Radial Basis Function, RBF, [Powell, 1992]), se basan en la distancia (usualmente euclídea) entre el vector de entrada y un vector llamado Prototipo o Centroide [Bishop, 1995]. Estos esquemas pueden ser entrenados mediante métodos mucho más rápidos que los utilizados para entrenar las redes basadas en el perceptrón multicapa.

Supóngase el "mapeado" de vectores $\mathbf{x}$ de un espacio de dimensión n a elementos t de un espacio unidimensional. Se dispone de N vectores $\mathbf{x}^n$ y de sus correspondencias t^n. Se desea encontrar una función que verifique

$$h(\mathbf{x}^n) = t^n \qquad n = 1, \cdots, N. \tag{2-32}$$

La solución RBF a este problema general es el uso de una serie de N funciones, una por dato, que tengan la forma $\phi(\|\mathbf{x} - \mathbf{x}^n\|)$ donde $\phi(\cdot)$ es una función no lineal. Se toma como función de mapeado una combinación lineal de estas funciones:

$$h(\mathbf{x}) = \sum_n w_n \phi(\|\mathbf{x} - \mathbf{x}^n\|) \tag{2-33}$$

y la ecuación (2-32) puede escribirse según la forma

$$\Phi \mathbf{w} = \mathbf{t} \tag{2-34}$$

donde $\mathbf{t}=\{t^n\}$, $\mathbf{w}=\{w^n\}$ y Φ es una matriz cuadrada cuyos elementos tienen la forma $\Phi^{nm} = \phi(\|\mathbf{x}^n - \mathbf{x}^m\|)$. Con los pesos $\mathbf{w}$ adecuados, la función $h(\mathbf{x})$ es una superficie continuamente derivable que pasa por los puntos $\mathbf{x}^n$. Si existe la matriz inversa, puede resolverse la ecuación respecto de $\mathbf{w}$ con

$$\mathbf{w} = \Phi^{-1}\mathbf{t}. \tag{2-35}$$

La función no lineal $\phi(\cdot)$ más utilizada es la gaussiana $\phi(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right)$, aunque también puede utilizarse la función:

$$\phi(x) = \left(x^2 + \sigma^2\right)^{-\alpha}, \quad \alpha > 0$$

Sin embargo, este método puede no funcionar correctamente en clasificación si hay ruido en los datos. El ruido en los datos puede hacer que la interpolación sea una superficie que oscile alrededor de la superficie óptima. Una solución a esto es hacer que la red contenga un número de puntos de interpolación mucho menor al número de datos. Además, el parámetro σ será diferente para cada una de las funciones y su valor se obtendrá a partir del entrenamiento. Por último, se añadirá un parámetro de sesgo a la suma (2-33) para compensar las diferencias entre la media de los datos de entrada y la media deseada en la salida. El esquema se ilustra en la Figura 2-20.

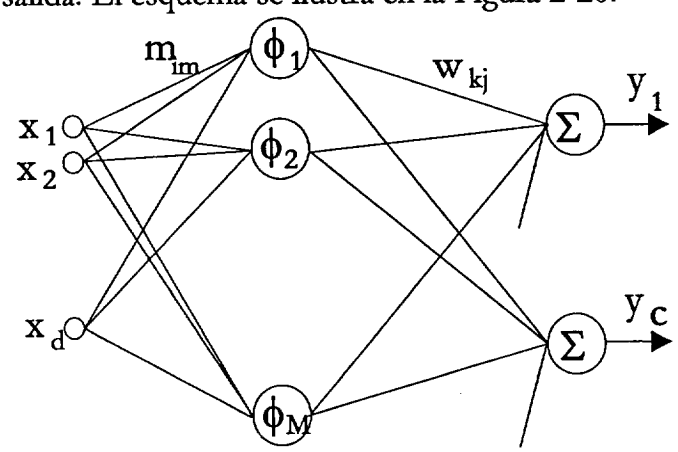


Figura 2-20. Esquema de una red de funciones de base radial

Si se utilizan gaussianas, cada una de las funciones de base radial de esta red tiene la forma

$$\phi(x) = \exp\left(-\frac{\|x - c_i\|^2}{2\sigma_i^2}\right) \qquad (2\text{-}36)$$

El entrenamiento de esta red puede hacerse de manera supervisada mediante el cómputo del gradiente del error cuadrático entre la salida obtenida y la deseada respecto de sus parámetros.

2.4.4.c Mapas autoorganizativos

Una manera de situar de forma no supervisada los centros de las funciones de base radial es el mapa autoorganizativo de Kohonen [Kohonen, 1990] . En este algoritmo los datos son introducidos en un conjunto de neuronas cada una de las cuales representa un centroide que apunta a una coordenada de una malla bidimensional. Como en 2.4.4.b, el vector de cada una de las neuronas,

es comparado con el dato mediante la distancia euclídea entre éste y cada uno de los centroides.

La adaptación se hace de la siguiente manera: primero se busca la neurona cuyo centroide $\mathbf{c}_i$ sea el más cercano (tenga mínima distancia euclídea) respecto del vector de entrada (al cual llamaremos $\mathbf{x}$). Después se actualizan todos los centroides que estén en la vecindad de $\mathbf{x}$. La actualización se hace de manera que el vector se mueva ligeramente hacia la posición del vector de entrada, es decir:

$$\mathbf{c}_j[n+1] = (1-\alpha[n])\mathbf{c}_j[n] + \alpha[n]\mathbf{x}[n] \qquad (2\text{-}37)$$

El valor de $\alpha[n]$ depende de la distancia entre el centroide $\mathbf{c}_i$ y el dato $\mathbf{x}[n]$ y del tiempo. $\alpha[n]$ puede ser binaria (1, 0) (según si el dato está o no más cerca del centroide que un umbral) o también una función radial decreciente con valor máximo igual a 1 que dependa de la distancia entre $\mathbf{c}_i$ y $\mathbf{x}$. Por ejemplo

$$\alpha[n] = \alpha_0 \exp\left(-\|\mathbf{x}-\mathbf{c}_i\|^2/\sigma^2[n]\right)$$

con $\sigma[n]$, en general, decreciente con el tiempo.

Si se utiliza este esquema como clasificador de patrones, donde las señales se deben agrupar en clases, el problema se convierte en un problema de decisión. En este caso se asigna un centroide de un determinado conjunto a cada una de las clases a clasificar. El dato de entrada queda clasificado entonces por la etiqueta del centroide al cual está más próximo. Al principio, deberá dejarse el algoritmo actuar según la regla de aprendizaje de la ecuación (2-37). Al final de la convergencia, deberá determinarse cuál es la etiqueta de cada uno de los centroides introduciendo una serie de datos de entrada cuya clase sea conocida.

Para aumentar la precisión de la clasificación se puede utilizar el algoritmo siguiente después del (2-37) (Cuantificación de Vectores de Aprendizaje, Tipo 1):

$$\begin{aligned}\mathbf{c}_i[n+1] &= (1-\alpha)\mathbf{c}_i[n] + \alpha\mathbf{x}[n], \text{si } \mathbf{x}[n] \text{ se ha clasificado correctamente}\\ \mathbf{c}_i[n+1] &= (1+\alpha)\mathbf{c}_i[n] - \alpha\mathbf{x}[n], \text{si } \mathbf{x}[n] \text{ se ha clasificado incorrectamente}\end{aligned} \qquad (2\text{-}38)$$

El parámetro α tendría la misma forma que en la regla anterior.

Un segundo método tendría la siguiente forma (Cuantificación de Vectores de Aprendizaje, Tipo 2): se define una ventana de anchura determinada que caiga en la intersección entre cada dos clases. Se considera que el dato cae en la ventana entre dos centroides $\mathbf{c}_i$ y $\mathbf{c}_j$ si las distancias d_i y d_j del dato con respecto a los centroides cumple

$$\min\left[d_i/d_j, d_j/d_i\right] > (1-\omega)/(1+\omega)$$

Si el vector de entrada cae en esa ventana y además la clasificación es errónea, se efectúa una corrección de tipo "antirreforzamiento", es decir:

$$\begin{aligned} \mathbf{c}_i[n+1] &= (1-\alpha)\mathbf{c}_i[n] + \alpha \mathbf{x}[n] \\ \mathbf{c}_j[n+1] &= (1+\alpha)\mathbf{c}_j[n] - \alpha \mathbf{x}[n] \end{aligned} \quad . \tag{2-39}$$

donde $\mathbf{c}_i$ es el centroide más cercano pero $\mathbf{x}[n]$ pertenece a $\mathbf{c}_j$ y además el dato está dentro de la ventana. El parámetro ω debe valer entre 0,1 y 0,2.

Si la clasificación se ha efectuado correctamente o el vector no ha caído en la ventana, no se hace nada con los centroides. Si se hace esto, las correcciones moverán los centroides, y con ellos la ventana, hasta que ésta se sitúe en la intersección de las distribuciones de probabilidad de las clases.

Sin embargo, este tipo de entrenamiento tiene el fallo de que, siendo la corrección de la clase correcta ($\mathbf{c}_j$) mayor que la de la clase incorrecta ($\mathbf{c}_i$), al ser ambas proporcionales a la distancia euclídea, las dos clases pueden acercarse sucesivamente con los pasos de entrenamiento. Además, puede caerse en situaciones de equilibrio subóptimas (mínimos locales).

Para evitar lo anterior se puede llevar a cabo el siguiente entrenamiento (Cuantificación de Vectores de Aprendizaje, Tipo 3)

$$\begin{aligned} \mathbf{c}_i[n+1] &= (1+\alpha)\mathbf{c}_i[n] - \alpha \mathbf{x}[n] \\ \mathbf{c}_j[n+1] &= (1-\alpha)\mathbf{c}_j[n] + \alpha \mathbf{x}[n] \end{aligned} \tag{2-40}$$

si $\mathbf{c}_i$ es el centroide más cercano pero $\mathbf{x}[n]$ pertenece a $\mathbf{c}_j$ y además el dato está dentro de la ventana.

$$\mathbf{c}_k[n+1] = (1+\varepsilon\alpha)\mathbf{c}_k[n] - \varepsilon\alpha \mathbf{x}[n], \tag{2-41}$$

para $k = \{i,j\}$, y $\mathbf{x}$, $\mathbf{c}_i$, $\mathbf{c}_j$ pertenecen a la misma clase.

El valor de ε es menor que 1 y depende de la anchura de la ventana, siendo menor para anchuras menores de ésta.

2.4.5. Alternativas a la igualación no lineal directa

Con un diseño adecuado al problema de la igualación es posible desarrollar aproximaciones que utilicen subsistemas lineales o de pequeño orden. Para ello, es necesario que cada uno de estos subsistemas se especialice en una parte del espacio de datos. Una de las aproximaciones existentes se basa en las redes de Jordan: es el esquema modular. Una propuesta alternativa, que puede resolver alguno de los inconvenientes de esta red es el esquema en escalera.

2.4.5.a Esquemas modulares

Este método se basa en el entrenamiento simultáneo de varias estructuras (lineales o no), cuya salida se suma ponderadamente para proceder a la decisión del símbolo. Cada una de las ponderaciones se corresponde con la probabilidad estimada de la verosimilitud de la salida de cada una de las redes. Estas ponderaciones las calcula un sistema que funciona paralelamente a las redes, y lo

hace basándose asimismo en los datos de entrada.

El entrenamiento de la red se hace mediante un algoritmo de gradiente. Para ello se utiliza la función de coste introducida por Jordan:

$$C = -\log\left(\sum_{i=0}^{N-1} p_i \exp\left(-\frac{1}{2}(x - y_i)^2 \right) \right) \qquad (2\text{-}42)$$

siendo x la señal deseada, y_i la señal de cada uno de los expertos y p_i la salida del supervisor. Se debe expresar el gradiente del coste en función de cada uno de los vectores de pesos de las redes, y actualizarlos según la dirección de máxima variación del coste, como es habitual.

La diferencia de este método respecto de los anteriores es que éste permite que cada una de las redes o módulos se especialice en una parte del espacio muestral. Además, en igualación [Cid, 1994] puede aplicarse la estructura lineal que proponen los autores para cada uno de los expertos, con buenos resultados.

El inconveniente de este algoritmo es que el hecho de que cada uno de los expertos esté aplicado a una parte del espacio hace que sus pesos sólo se actualicen cuando la verosimilitud de su resultado sea alta, es decir, cuando el peso p_i que estima la verosimilitud de la salida i sea cercano a 1. Esto ralentizará el entrenamiento respecto de redes que se actualicen con todos los datos. En efecto, el gradiente del coste (2-42) respecto de la salida i será

$$\frac{\partial C}{\partial y_i} = q_i (x + y_i),$$

$$q_i = \frac{p_i \exp\left(-\frac{1}{2}(x - y_i) \right)}{\sum_{j=0}^{N-1} p_j \exp\left(-\frac{1}{2}(x - y_j) \right)} \qquad (2\text{-}43)$$

Este algoritmo, sin embargo, es ventajoso respecto del perceptrón multicapa precisamente porque cada elemento de su estructura se adapta a una parte del espacio muestral, mientras que aquélla hace que todos los pesos se adapten simultáneamente a todo el espacio, lo que acentúa el conocido problema de la caída en mínimos locales.

2.4.5.b Esquema en escalera

Esta aproximación se plantea, inicialmente [Figueiras/1, 1996] [Figueiras/2, 1996)], como solución al problema de aplicar la función de coste de

Kullback-Leibler[12] [Haykin, 1994].

Considérese el problema de clasificación del conjunto r de elementos en las clases $1 \leq j \leq N$ mediante funciones de activación $o_j(w)$. Esta salida puede interpretarse como una estimación de la probabilidad a posteriori de que la hipótesis j sea cierta, es decir:

$$o_j(w) = \hat{P}(j \mid r) \tag{2-44}$$

Llamemos $P(j \mid r)$ a la probabilidad real a posteriori sobre los símbolos, es decir, la probabilidad de que la hipótesis j sea cierta a la vista de la entrada r. Sea $P(r)$ la distribución de probabilidad a priori de los datos. Puede definirse la entropía relativa para la salida del clasificador de la siguiente forma:

$$C(w) = E\{H(r)\} = \int P(r)\left\{\sum_i o_j \ln\frac{o_j}{P(j \mid r)} + (1-o_j)\ln\frac{1-o_j}{1-P(j \mid r)}\right\}dr \tag{2-45}$$

Esta función se basa en la medida de la información contenida en una determinada observación, con respecto a la hipótesis de decisión. En otras palabras, esta función es una estimación de la entropía de la entrada r relativa a la de la hipótesis.

La aplicación directa de esta función en esquemas FIR de recepción no binarios no es posible, ya que la salida no está acotada entre –1 y 1. Si se aplica una función de saturación a la salida del filtro se destruye la linealidad anterior al decisor multinivel: al no haber proporcionalidad entre los niveles transmitidos y las entradas al cuantificador multinivel, no es posible clasificar correctamente las muestras.

Para poder aplicar una función de coste del tipo descrito en (2-45) es necesario que el esquema de recepción sea un esquema neuronal. La red neuronal debe tener un número de salidas igual a la mitad del número de niveles del alfabeto utilizado para cada una de las dimensiones de la constelación utilizada. Cada una de las salidas debe estar acotada entre –1 y 1. A tal efecto se usa la activación Softmax [Haykin, 1994], que normaliza cada una de las salidas respecto de la suma de todas ellas. La expresión para la salida i es:

$$o_i = \frac{\exp\left(\mathbf{w}_i^{L\,T}\mathbf{x}\right)}{\sum_{j=1}^{K}\exp\left(\mathbf{w}_j^{L\,T}\mathbf{x}\right)} \tag{2-46}$$

donde $\mathbf{w}_j^{L\,T}$ es el vector de pesos de la neurona j de la última de las L capas de la red, y x es el vector de entrada a esta neurona (véase notación en 2.4.4.a).

---

[12] Véase la sección 3.4.2 para una exposición más detallada de este criterio.

Como se expone en 2.4.4, la aplicación de una red multicapa presenta el inconveniente de la lentitud en la convergencia, además del riesgo de la caída en mínimos locales.

El algoritmo en escalera se lleva a cabo como se explica a continuación. A efectos de la representación gráfica, pondremos como ejemplo un canal $H(z) = 1 + 0.2z^{-1}$ y una señal 8-PAM. Se toman vectores de dos muestras para hacer la igualación. Supóngase que los coeficientes de los filtros están adecuadamente calculados para discriminar señales. El símbolo recibido en el instante actual y las rectas de discriminación definidas por la ecuación (2-30) para cada uno de los filtros se representan en la Figura 2-21

- Primero se decide con la recta 0. El resultado es negativo, con lo cual se pone signo positivo a los términos independientes, y las rectas de discriminación se sitúan en la parte negativa. La señal está entre el subconjunto {-1, -3, -5, -7}.
- Se decide un resultado positivo con la recta 2, con lo cual la señal está en la parte alta, es decir, {-1, -3}.
- Como la señal está en la parte alta, se utiliza la recta 1 para decidir que la señal vale -1

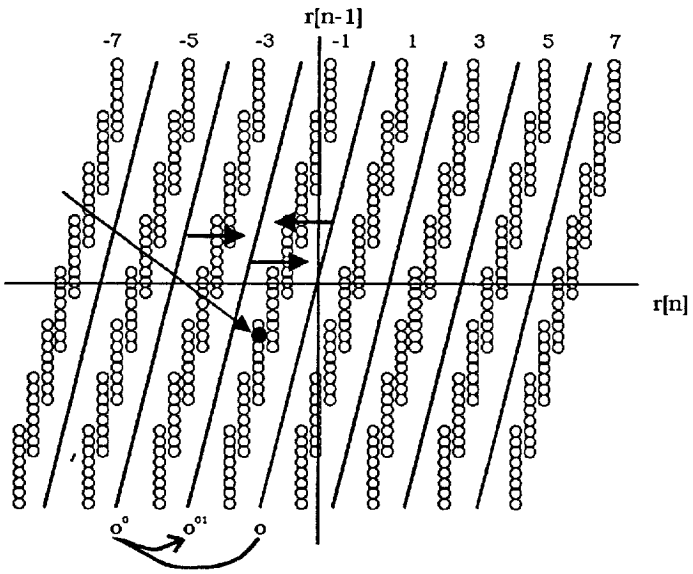


Figura 2-21. Secuencia de decisiones del algoritmo en escalera para el caso 8-PAM y espacio de datos de dimensión 2. Canal fase mínima tipo $H(z)=1+0.2z^{-1}$.

El aprendizaje de este sistema puede ser el habitual del filtro FIR adaptativo para cada uno de los filtros, con la salvedad de que cada señal deseada debe expresarse en forma de una secuencia determinada de decisiones (o, lo que es lo mismo, como una secuencia binaria). Mediante la oportuna expresión del filtro en forma compleja se puede hacer extensivo a señales QAM, aunque aquí nos limitamos al estudio del comportamiento del filtro en su versión para señales

reales.

Este esquema abre algunas posibilidades, que serán objeto de estudio en los siguientes capítulos. Por un lado, respecto de la forma no lineal del canal. Tal como se ha visto en el apartado 0, las no linealidades que se pueden encontrar en un canal de comunicaciones se manifiestan en forma de funciones de compresión. En algunos casos es suficiente con utilizar un único filtro FIR con salida lineal y que el decisor duro multinivel tenga niveles de decisión adaptativos. Pero en otros casos es necesario utilizar fronteras de decisión diferenciadas, ya que las fronteras óptimas no serán paralelas. Aquí es preciso utilizar filtros FIR especializados en subconjuntos de símbolos, de manera que la aproximación sea mejor. El igualador en escalera cumple esta función: todos los filtros están especializados en algún subconjunto de símbolos, y los filtros terminales están especializados en sólo una pareja de ellos. Por ello, si hay M símbolos, se necesitan M/2 decisiones posibles y, por simetría, M/4 filtros terminales.

Otro aspecto respecto de la no linealidad del canal es que ésta hace que la máxima probabilidad de error se produzca en los bordes del conjunto de datos, mientras que en el interior, donde las propiedades son aproximadamente lineales, la probabilidad de error se mantiene en niveles mucho más reducidos. Ello sugiere la posibilidad de utilizar modelos no lineales de igualación en los extremos, mientras que se puede utilizar una aproximación lineal fuera de ellos. Esto es posible dado que se usan filtros separados: los filtros de la parte exterior pueden ser enteramente o parcialmente (a tramos) no lineales. Con esto existe la posibilidad de llegar a tener las prestaciones de un filtro no lineal reduciendo la complejidad de cálculo.

ANEXO 2.1 SEPARABILIDAD DE CLASES EN SEÑALES PAM

A2.1.1 Señal PAM en canal lineal

Sea un canal $H(z)=1+h_1z^{-1}$, (es decir, $\mathbf{h}=[1\ h_1]^T$), emisión PAM de P niveles y un receptor que toma vectores de muestras de orden 2. Sean las clase 1 y 2 correspondientes a

$$\mathbf{r}_1 = \begin{bmatrix} x_1 & x_j \\ x_j & x_k \end{bmatrix}\mathbf{h}$$
$$\mathbf{r}_2 = \begin{bmatrix} x_2 & x_i \\ x_i & x_m \end{bmatrix}\mathbf{h} \qquad (2\text{-}47)$$

Dos vectores $\mathbf{r}_1$, $\mathbf{r}_2$ correspondientes a dos clases contiguas no serán separables si su vector distancia es nulo. Si se calcula la distancia de estos vectores a partir de la ecuación (2-47) teniendo en cuenta que $x_{i+1}-x_i=\Delta$ y se iguala a cero se obtiene

$$\mathbf{r}_2 - \mathbf{r}_1 = \begin{bmatrix} \Delta & x_i - x_j \\ x_i - x_j & x_m - x_k \end{bmatrix}\mathbf{h} = \mathbf{0} \qquad (2\text{-}48)$$

De la primera fila se deduce

$$x_j - x_i = \frac{\Delta}{h_1}$$

y a partir de esto se halla la condición de separabilidad lineal: de la segunda fila se despeja h_1. El valor más pequeño de h_1 se dará cuando $x_m-x_k=2x_{max}=(P-1)\Delta$ (siendo P el número de símbolos) y su valor será:

$$h_1 < \frac{1}{\sqrt{(P-1)}} < 1 \qquad (2\text{-}49)$$

lo que asegura la separabilidad de las clases.

Para que las clases sean linealmente separables, la clasificación debe poder hacerse con discriminadores lineales que tengan la forma (1, w), es decir, los conjuntos serán linealmente separables si las distancias $d_1=\mathbf{r}_1\mathbf{w}^H$ y $d_2=\mathbf{r}_2\mathbf{w}^H$, con $\mathbf{w}=\{1, w\}$ verifican

$$d_1 = \mathbf{r}_1\mathbf{w}^H > k$$
$$d_2 = \mathbf{r}_2\mathbf{w}^H < k$$

En tal caso,

$$d_1 < d_2$$

Desarrollando esta desigualdad para los vectores de la ecuación (2-47) y teniendo en cuenta que $x_i-x_{i+1}=\Delta$, se obtiene:

$$d(w) = \Delta + h_1(x_i - x_j) + w(x_i - x_j) + h_1 w(x_m - x_k) > 0 \qquad (2\text{-}50)$$

Se escoge

$$-w = h_1 < \frac{1}{\sqrt{2(P-1)}} \qquad (2\text{-}51)$$

valor que maximizará el término a la izquierda de la desigualdad[13]. En efecto, si se reescribe la ecuación (2-50):

$$d(w) = \Delta + h_1 z_1 + w z_1 + h_1 w z_2$$

se puede escribir

$$d(w = -h_1) = \Delta - h^2 z_2$$

$$d(w = -h_1 + \varepsilon) = \Delta - h^2 z_2 + \varepsilon z_1 + \varepsilon h z_2$$

y se concluye que la ecuación d(·) es mínima para $\varepsilon=0$, ya que para cualquier otro valor de ε y para algún par de valores z_1 y z_2 escogidos adecuadamente entre los extremos $-2x_{max} \le z_1, z_2 \le 2x_{max}$, d(·) es mayor que $d(w=h_1)=1-h^2_1 z_2$. Por tanto, (2-49) y (2-51) son las condiciones de separabilidad para el caso de transmisión 8-PAM con canal lineal.

A2.1.2 Señal PAM en canal con no linealidad sin memoria

Si se repite el cálculo para no linealidad F(·) en la entrada del canal (no linealidad sin memoria), se obtiene que hay separabilidad lineal si:

$$h_1 < \frac{\sqrt{F(x_P) - F(x_{P-1})}}{\sqrt{2F(x_P)}} \qquad (2\text{-}52)$$

$$w = -h_1$$

x_P es el símbolo de mayor amplitud.

Este valor de h_1 es menor que en canal lineal. Esto puede verse inmediatamente teniendo en cuenta que F(·) tiene derivada primera positiva y monótona decreciente: la no linealidad de la entrada disminuye el rango de valores de h_1 o amplitud máxima de la ISI para que haya separabilidad con una amplitud dada de la señal PAM. En cualquier caso, si las clases son separables, o

---

[13] La solución de Wiener es $w=h_1/(1-h_1)$, lo que minimiza el error cuadrático medio; es decir, para este caso en particular:

$$E[\{h_1 x_j + h_1 w x_k + w x_k\}^2] = h_1^2 E\{x_j^2\} + h_1 w^2 E\{x_k^2\} = 0 \Rightarrow w = h_1/(1-h_1),$$

pero con la perspectiva con la que discutimos este caso, el criterio consiste en la maximización de la distancia entre dos clases adyacentes.

lo que es lo mismo, la distancia entre cualesquiera dos centroides de la clase x_{i+1} y la clase x_i es diferente de cero, serán separables linealmente.

Es importante resaltar que el valor óptimo hallado para w en la ecuación (2-50) es válido para cualquiera de las rectas de discriminación: es decir, las rectas de discriminación serán paralelas.

A2.1.3 Señal PAM con no linealidad con memoria

Si se tiene un canal con no linealidad G(·) a la salida (con memoria) entonces la señal presente a la entrada del receptor tendrá la forma

$$\mathbf{r}_1 = \begin{bmatrix} G\left(x_1 + h_1 x_j\right) \\ G\left(x_j + h_1 x_k\right) \end{bmatrix}$$

$$\mathbf{r}_2 = \begin{bmatrix} G\left(x_2 + h_1 x_1\right) \\ G\left(x_1 + h_1 x_m\right) \end{bmatrix} \qquad (2\text{-}53)$$

El solapamiento se producirá cuando para algún par de centroides de clases diferentes la distancia sea cero. Al ser la función G(·) monótona creciente y, por lo tanto, biyectiva, es fácil ver que el solapamiento se producirá si h_1 sobrepasa el límite:

$$h_1 < \frac{1}{\sqrt{(P-1)}} < 1 \qquad (2\text{-}54)$$

Es decir, si no hay solapamiento de clases antes de la no linealidad, tampoco la habrá después. Esta es la diferencia con el anterior tipo de saturación. En el caso en que haya separabilidad lineal en canal lineal (en ausencia de ruido), y teniendo en cuenta que G(·) es monótona creciente, también habrá separabilidad lineal cuando esta no linealidad esté presente a la salida del canal.

Además, como la decisión final es llevada a cabo por diferentes decisores que dependen de la posición del dato en el espacio, estos decisores podrán ser lineales o no lineales dependiendo de si la amplitud del dato es baja (la distorsión será poco apreciable) o alta (la distorsión afectará más al dato y por lo tanto será ventajoso aplicar esquemas no lineales en esa zona del espacio).

3. ALGORITMO EN ESCALERA

3.1. DESCRIPCIÓN DEL ALGORITMO

El esquema de funcionamiento del algoritmo en escalera se basa en una cadena de decisiones binarias, tal como hemos visto en el apartado 2.4.5.b. En esta sección se detalla la estructura en escalera y se introduce la notación a seguir. En la sección 0 se explica el entrenamiento del esquema en escalera, y se presentan diversas funciones de coste para entrenamiento por minimización de gradiente. La sección 0 está dedicada a ensayos del algoritmo para algunos tipos de canal no lineal que se han considerado significativos.

3.1.1. Clasificación en escalera

Las comunicaciones digitales utilizan símbolos que se codifican en el emisor a partir de conjuntos de L bits tomados de la fuente de información. Por ello los símbolos se pueden clasificar en árbol de la siguiente manera (Figura 3-1): dado un conjunto X de símbolos, en el primer escalón, la clasificación del conjunto se divide en dos partes: X^0 es el subconjunto de elementos cuyo primer bit (bit de signo) es cero; X^1 es el subconjunto de elementos cuyo primer bit es uno. En el segundo escalón habrá dos clases, que pueden subdividirse a su vez de la misma manera: cada uno de los dos subconjuntos de X^{b_0}, $b_0=\{0, 1\}$ se divide

en dos subconjuntos $X^{b_0 b_1}$ $b_1=\{0, 1\}$ dependiendo del valor del segundo bit. Para el escalón n habrá 2^n clases $X^{b_0 \cdots b_{n-1}}$ que se subdividirán en 2^{n+1} en función de los valores de los bits.

3.1.2. Discusión

La ventaja de la partición del espacio mediante clasificadores binarios frente al clasificador multinivel es que cada uno de los clasificadores binarios sólo se adapta a una parte del espacio de datos: aquella parte en la que están las dos clases que debe separar (adaptación local), mientras que el clasificador multinivel debe estar adaptado a todo el espacio de datos (adaptación global). No es necesario decir mucho más para comprender que el clasificador basado en la partición producirá mejores resultados que el multinivel en aquellos casos en que las características del espacio no sean lineales.

Al utilizar varios clasificadores, en principio independientes, unos pueden tener distinta estructura que otros: pueden utilizarse esquemas basados en funciones lineales en zonas en las que haya una buena separabilidad lineal y esquemas basados en funciones no lineales en aquellas zonas en las que no la haya.

Por otra parte, si se usa un esquema de clasificadores binarios, la decisión necesita, en principio, la observación de todas y cada una de las salidas de los decisores. Así, si se tiene un alfabeto de P símbolos, hay que hacer L-1 observaciones para llevar a cabo la decisión.

Si se utiliza un algoritmo de decisión en escalera, el número de observaciones por dato recibido se reduce. En efecto, si se tiene un alfabeto de P

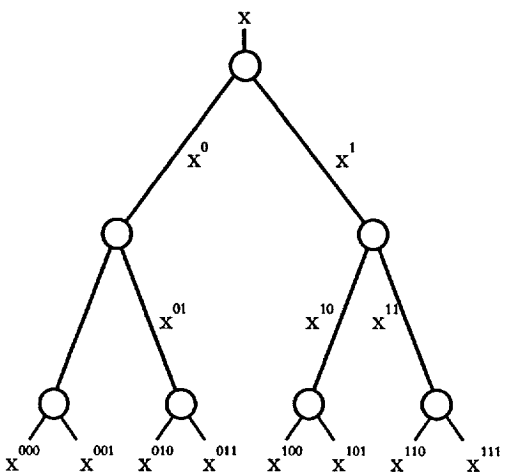


Figura 3-1. Clasificación en escalera y notación adoptada.

símbolos formados con L bits, el número de clases es 2^L. El número de clasificadores para el escalón n es 2^{n-1}. Suponiendo que hay simetría en el espacio de datos, el número total de clasificadores es $\sum_{n=0}^{L-1} 2^n = 2^L - 1$. Una ventaja de la clasificación en escalera es que deben realizarse L decisiones, una por escalón, para cada uno de los símbolos, mientras que si se emplea una partición directa del espacio de datos, para cada símbolo deben examinarse todos los 2^L-1 clasificadores. Esto supone un importante ahorro de coste computacional.

Puede objetarse, a primera vista, que cada uno de los elementos del clasificador en escalera no contiene sólo las características locales del espacio como sí las contiene cada uno de los clasificadores de una partición directa: el primero de los igualadores utiliza todos los datos, los que están en su entorno y los que están lejos. Sin embargo, este clasificador sigue siendo binario y no multinivel: se puede pensar que este clasificador utiliza información grosera o poco refinada de todo el espacio, y que los sucesivos clasificadores refinan esta información. La diferencia entre el primer clasificador binario del esquema en escalera y un igualador multinivel es que el primero no necesita emplear información precisa acerca de cada una de las clases que separa. Por supuesto, lo mismo reza para los demás elementos de la estructura en escalera.

Por último, el uso de clasificadores binarios permite utilizar en forma directa funciones de coste cuya entrada sea la salida del filtro pasada por una función de tipo tangente hiperbólica, mientras que no está claro cómo utilizar este tipo de activación en decisión multinivel. En esta misma línea, es posible utilizar algoritmos de clasificación binaria para los que una extensión a clasificación multinivel no es inmediata. Entre ellos destacamos la Máquina de Vectores de Soporte (SVM) [Vapnik, 1995] [Schölkopf, 1997] [Burgues, 1997]. En el capítulo 5 se propone y examina una solución inspirada en la idea del clasificador SVM.

3.1.3. Igualadores en escalera

Se tiene un conjunto de clasificadores (que supondremos aquí perfectamente adaptados para minimizar el error cuadrático medio de su salida con respecto al símbolo emitido). El primer clasificador (o) separa las señales en dos clases dependiendo del valor de su bit más significativo, es decir, dependiendo del valor de su signo. Cada una de las anteriores clases es dividida por sendos clasificadores (salida o^{b_0}) en función del valor del segundo bit más significativo. Para cada una de las clases de este clasificador habrá otro (salida

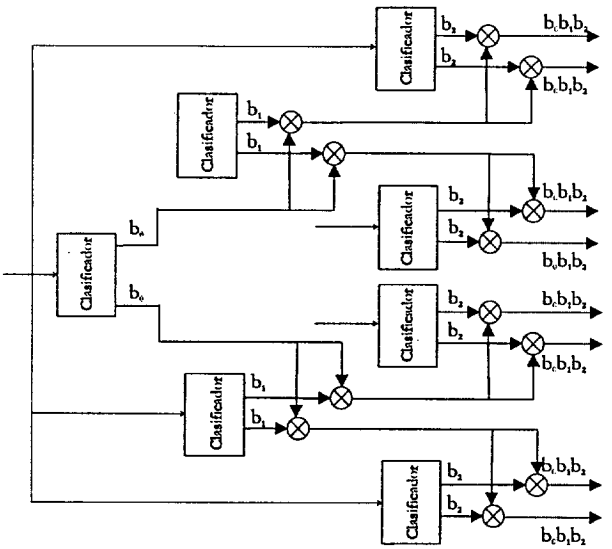


Figura 3-2. Igualador en escalera para recepción 8-PAM.

$o^{b_0 b_1}$) que la dividirá en dos dependiendo del valor del tercer bit más significativo, y así sucesivamente hasta llegar al bit menos significativo.

Cuando una señal r[n] es aplicada a la entrada del primer clasificador, su salida se utiliza no sólo para determinar el valor del bit más significativo, sino para escoger cuál de los dos clasificadores o^{b_0}, $b_0=\{0,1\}$, se utilizará en la siguiente decisión. Lo mismo vale para la decisión de cada uno de los clasificadores escogidos en los sucesivos pasos del algoritmo. El conjunto de las L decisiones (signo de $o\, o^{b_0}\, o^{b_0 b_1}$ ...) da el valor del símbolo en binario.

La Figura 3-2 muestra el esquema del igualador en escalera. Visto de esta manera, el igualador es una cadena de decisiones "bit a bit". Cada clasificador tiene una salida binaria y su complementario. El primer clasificador decide el bit de signo b_0. La segunda capa, con dos clasificadores, decide el bit b_1, el resultado del cual se multiplica por el resultado del bit anterior, y así sucesivamente, las capas siguientes decidirán cada una un bit cada vez menos significativo hasta llegar al último. Para el caso de señales P-PAM, son necesarios L bits, donde $P=2^L$ y, por lo tanto, L capas. El número de clasificadores necesarios es $P-1=2^L-1$. Sólo habrá una salida no nula, que corresponderá a la decisión sobre el símbolo emitido.

En la Figura 3-3 se ve un ejemplo de secuencia de decisiones en una constelación 8-PAM para detectar el símbolo –1, correspondiente a la secuencia de bits 011. El clasificador central tiene una salida de signo negativo (bit más significativo igual a cero), por lo que se selecciona el clasificador o^0 para la

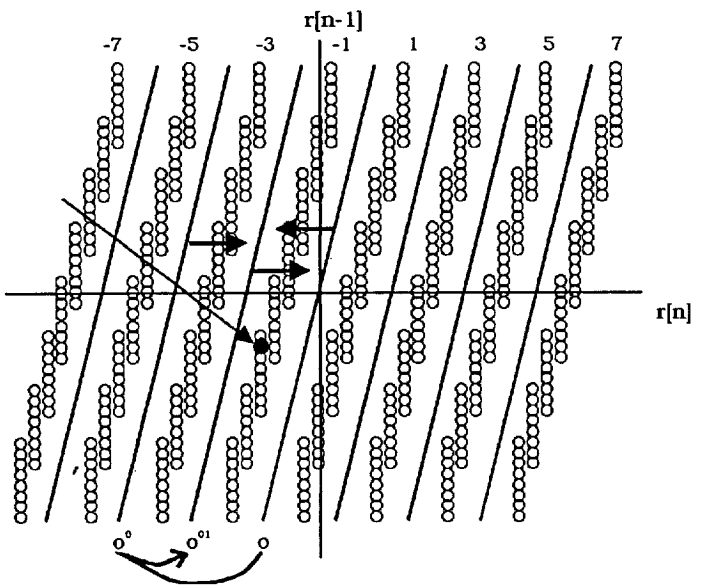


Figura 3-3. Secuencia de decisiones en el algoritmo en escalera para detectar el símbolo –1 (bits 011) en una constelación 8-PAM. El primer clasificador (o) decide que el primer bit es 0, con lo que se escoge como segundo clasificador o^0, que determina que el segundo bit es 1. Entonces se selecciona o^{01}, cuya salida produce la decisión 0 para el tercer y último bit.

siguiente decisión. Este clasificador tiene salida positiva (segundo bit igual a 1), lo que hace que se seleccione el clasificador o^{01}. Éste tiene salida positiva (tercer bit igual a 1), lo que finaliza la secuencia de decisiones.

Por otra parte, la mitad de los decisores de cada una de las capas puede ser obviado cuando en el espacio de datos haya simetría. Para un canal lineal de la forma $H[z]=\sum_{k=0}^{N-1} h_k z^{-k}$ y un receptor que toma vectores de M muestras consecutivas, los datos de entrada al igualador (en ausencia de ruido) tienen la forma

$$\mathbf{r}[n]=\begin{Bmatrix} r[n] \\ \vdots \\ r[n-M+1] \end{Bmatrix}=\begin{Bmatrix} \sum_{k=0}^{N-1} h_k x[n-k] \\ \vdots \\ \sum_{k=0}^{N-1} h_k x[n-k-M+1] \end{Bmatrix}=\mathbf{X}[n]\mathbf{h} \qquad (3\text{-}1)$$

donde x[n] es el símbolo emitido en el instante n, y $\mathbf{X}[n]$ es una matriz cuyas M filas son los vectores $\mathbf{x}[n-k]=\{x[n-k], ..., x[n-k-N+1]\}$, $0 \le k \le M-1$

Si hay simetría en los datos, las fronteras de decisión a ambos lados del decisor central también serán simétricos, lo que significa que se puede utilizar un

clasificador de la parte positiva para la parte negativa tan sólo girándolo 180° respecto del origen. De esta manera sólo es necesario utilizar $2^{L-1} = P/2$ clasificadores.

Equivalentemente, se puede eliminar el signo de la señal de entrada y proceder en las siguientes decisiones como si estuviera situado en la parte del espacio de datos correspondiente a los símbolos positivos. Después del proceso, se restaura el signo. Si se procede de esta manera, el esquema equivalente es el de la Figura 3-4.

3.1.4. Esquema de entrenamiento del igualador en escalera

El esquema completo del igualador en escalera adaptativo está en la Figura

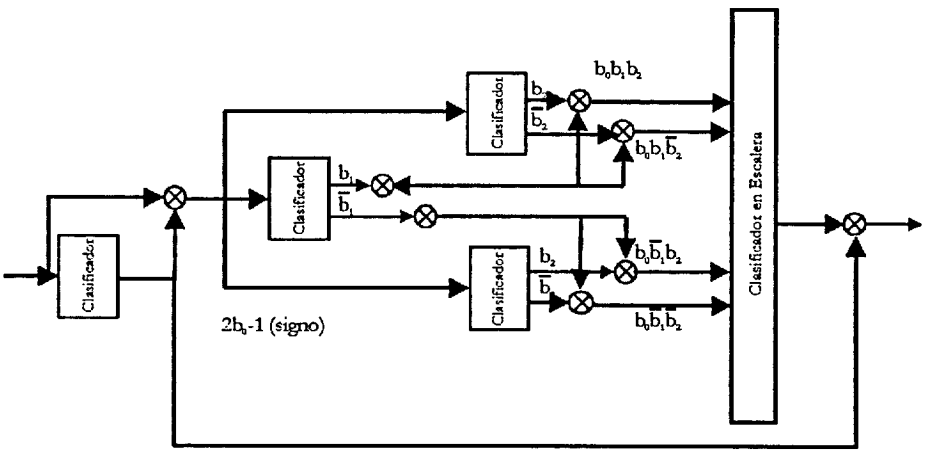


Figura 3-4. Simplificación del igualador aprovechando la simetría del conjunto de datos. Particularización a 8-PAM.

3-5. La entrada r[n] corresponde a la señal convolucionada por un filtro adaptado al pulso sobre el que se modulan los símbolos en el emisor y muestreada en los instantes kT_s, siendo T_s el periodo de símbolo.

Cada uno de los P/2 decisores necesarios está formado por un filtro FIR de M coeficientes $\mathbf{w}_m^{b_0 \cdots b_i}$ (más un término de "offset") más un decisor duro. La salida de cada uno de los filtros puede pasarse por una tangente hiperbólica para utilizar su salida en el cálculo del error. Se aplica el algoritmo en escalera al conjunto de decisiones.

El entrenamiento de los filtros se lleva a cabo inicialmente en forma

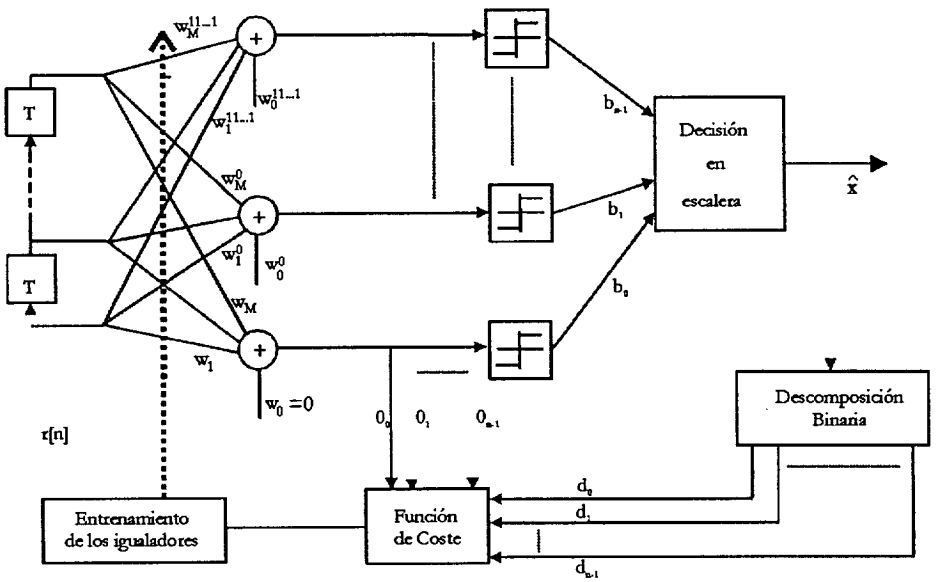


Figura 3-5. Estructura del igualador en escalera.

guiada: el receptor tiene conocimiento previo de los valores del primer conjunto de símbolos emitidos, información que utiliza para llevar los filtros a la situación de mínimo coste. Una vez recibido este conjunto de símbolos, se supone que el número de decisiones incorrectas es despreciable y por lo tanto puede utilizarse la decisión para calcular el error a la salida de los filtros (periodo de funcionamiento dirigido por decisión). Como casi todos los símbolos han sido correctamente detectados, el coste estimado en cada detección será casi siempre el real y debe ser, por lo tanto, válido para el posterior entrenamiento de los filtros, que será necesario para seguir las variaciones en la función de transferencia del canal.

En el caso de que los filtros sean polinómicos[14], la salida del clasificador $b_0...b_j$ será

$$o^{b_0\cdots b_j}\left[n\right]=\sum_{k=0}^{M-1} r\left[n-k\right] w_k^{b_0\cdots b_j} + w_M^{b_0\cdots b_j} \tag{3-2}$$

Si se define el vector $\mathbf{r}\left[n\right]=\left\{r\left[n\right],\cdots,r\left[n-M+1\right],1\right\}^T$ la salida es

[14] En el caso de filtros lineales, r[n] corresponde a las muestras de entrada al receptor y w_k son los pesos; para filtros polinómicos, r[n] corresponde a los productos cruzados del vector de entrada extendido (1, r'[n], r'[n-1], ...)^T y w_k son los núcleos de Volterra del filtro.

$$o^{b_0 \cdots b_j}[n] = w^{b_0 \cdots b_j\ T} r[n] \tag{3-3}$$

3.2. ENTRENAMIENTO DEL ALGORITMO EN ESCALERA

3.2.1. Entrenamiento de los filtros

El entrenamiento que se utilizará en los experimentos es del tipo descenso por la máxima pendiente. La actualización a emplear para los coeficientes de los filtros FIR es

$$w[k+1] = w[k] + \mu\left(-\nabla_{w(k)} E\left(\varepsilon_L^2\right)\right) \tag{3-4}$$

donde ε_L^2 es el error cuadrático de la salida respecto de la salida deseada.

Este método es conocido como método de máxima pendiente. El inconveniente es la imposibilidad de calcular de forma exacta este gradiente. Una forma de este algoritmo es el Least Mean Squares o LMS. Se basa en substituir la esperanza del error cuadrático por su valor instantáneo. Como se utiliza la función sin promediar, este método es ruidoso, pero este ruido se va promediando a lo largo de todo el proceso, actuando en este sentido como un filtro paso bajo. En lugar del error cuadrático, se pueden utilizar otros objetivos o funciones de coste.

3.2.2. Aplicabilidad de funciones de coste

Los algoritmos de decisión tienen salidas de decisión binaria. Esto permite introducir a la salida del filtro funciones de saturación, tales como la tangente hiperbólica, que limiten la salida a valores entre ±1. Introducir funciones de este tipo produce ciertas ventajas respecto de las salidas lineales.

Una de ellas es la disminución del ruido de desajuste del filtro con respecto a la solución de mínimo error medio. Aunque el filtro lineal entrenado con la función de coste error cuadrático medio converge en media al óptimo de Wiener bajo hipótesis no muy restrictivas (véase anexo 3.1), la desviación típica de la solución será alta para pasos de adaptación μ por encima de cierto valor. Hay que establecer, por tanto, un compromiso entre velocidad de convergencia y ruido de desajuste. Mediante el uso de la tangente hiperbólica se reduce el ruido

de desajuste, con lo que puede aumentarse μ para acelerar la convergencia.

Otra de las ventajas de las funciones de saturación a la salida de los filtros es la posibilidad de introducir la medida de la entropía relativa como función de coste, ya que sólo se puede utilizar esta función de coste si la salida del filtro está acotada.

Se han ensayado tres funciones de coste:

- función de coste error cuadrático medio con filtro de salida lineal;
- función de coste error cuadrático medio con salida tangente hiperbólica;
- función de coste medida de la entropía relativa o de Kullback-Leibler (que necesita la salida tangente hiperbólica).

En las siguientes secciones se describen brevemente las tres funciones de coste.

3.2.3. Función de coste error cuadrático

La primera de las funciones de coste empleada es la función error cuadrático, cuya expresión es

$$\varepsilon_L^2[n] = (d_n - o[n])^2 \tag{3-5}$$

siendo d_n la salida deseada para el filtro $\mathbf{w}$ en el instante n. Si la salida $o[n]$ es directamente la salida del filtro FIR llamamos a esta función, el gradiente de esta función respecto del vector de pesos del filtro es

$$\nabla_w \equiv 2(d_n - o[n])\mathbf{r}[n] \tag{3-6}$$

Si la salida tiene la forma

$$o[n] = \tanh(\lambda \mathbf{r}^T[n]\mathbf{w}) \tag{3-7}$$

(siendo λ real y positivo), el gradiente de ésta respecto del vector de pesos del filtro es

$$\nabla_w \equiv 2\lambda(d_n - o[n])(1 - o^2[n])\mathbf{r}[n] \tag{3-8}$$

3.2.4. Función de coste Entropía Relativa (Kullback-Leibler)

Cuando la salida del filtro se pasa a través de una función de saturación bipolar, es posible dar a su valor una interpretación probabilística. Definimos

$$\frac{1}{2}(1 - o) = P_0 \tag{3-9}$$

como la probabilidad estimada de que la decisión "-1" (o "0") sea cierta,

condicionada a la entrada r[n]. De igual forma podemos definir

$$\frac{1}{2}(1+o)=P_1 \tag{3-10}$$

como la probabilidad estimada de que la decisión "1" sea cierta condicionada a la entrada r[n]. Sean Q_0 y Q_1 las probabilidades (reales) condicionadas a r[n] de que las decisiones tomadas sean ciertas. Sean por último P(0) y P(1) las probabilidades a priori de los símbolos 0 y 1.

Se puede entonces definir la entropía relativa o medida de la información de Kulback-Leibler [Haykin, 1994] en los siguientes términos:

$$\begin{aligned} H &= P(0)Q_0 \ln\frac{Q_0}{P_0} + P(1)Q_1 \ln\frac{Q_1}{P_1} = \\ &= P(0)Q_0 \ln\frac{2Q_0}{1-o} + P(1)Q_1 \ln\frac{2Q_1}{1+o} \end{aligned} \tag{3-11}$$

Esta es una medida de la cercanía entre la probabilidad medida y la real. Cuando son iguales, la función se anula, y crece cuando la diferencia entre las dos probabilidades aumenta.

Las probabilidades Q_0 y Q_1 pueden substituirse por los valores deseados de la salida (convenientemente normalizados para que valgan 0 y 1). Siempre y cuando no haya peligro de caer en un mínimo local, las salidas o[n] de los filtros serán aproximadamente iguales a las probabilidades a posteriori de la distribución de los datos.

Si se sustituyen las probabilidades Q_0 y Q_1 por las salidas deseadas se obtiene:

$$H = (1+d[n])\ln\frac{1+d[n]}{1+o[n]} + (1-d[n])\ln\frac{1-d[n]}{1-o[n]} \tag{3-12}$$

Aunque esta ecuación necesita una gran cantidad de cómputo comparada con el coste cuadrático (un simple producto y una suma), su gradiente es muy sencillo: necesita menos cálculos que el gradiente del coste cuadrático cuando la salida utilizada es la tangente hiperbólica. Para este caso, el gradiente respecto de w es

$$\nabla_w \equiv 2\lambda(o[n]-d[n])r[n] \tag{3-13}$$

En lo sucesivo, el término constante 2 que aparece en los tres gradientes será obviado. No obstante se mantienen las proporciones para hacer comparaciones entre las tres funciones de coste.

3.2.5. Entrenamiento discriminativo

Uno de los objetivos del presente estudio de algoritmos de igualación es la búsqueda de esquemas de igualación no lineal que exijan el mínimo coste computacional posible. Aún utilizando filtros FIR, es posible igualar canales que presenten no linealidades, ya que cada uno de los filtros utilizados para construir los clasificadores actúa localmente. Si los símbolos de la constelación utilizada contienen L bits, entonces la decisión en el receptor involucrará la salida de L filtros. Pero, ¿es necesario entrenar los L filtros? La respuesta es no. Esto es debido a que el único filtro que recibirá una información importante para su convergencia es aquél cuya superficie de discriminación esté más cercana al dato presente en la entrada. Hay que resaltar dos situaciones:

- Cuando la salida del filtro pasa por una tangente hiperbólica, si el dato está muy alejado del filtro, la salida será prácticamente ±1, y el valor del gradiente de la función de coste será prácticamente nulo. Entrenar con este dato el filtro es inútil. Dicho de otra manera, entrenar todos los filtros no acelera la convergencia del algoritmo.
- Si la salida del filtro es lineal (y por tanto la función de coste es el error cuadrático) la situación es diferente. Cuando el dato está muy alejado del filtro la función de coste tendrá un valor muy alto y producirá desajuste: el entrenamiento se volverá ruidoso.

El entrenamiento no discriminativo corresponde al entrenamiento de los L filtros, mientras que el discriminativo consiste en entrenar sólo el filtro que esté más cerca del dato. La principal ventaja del entrenamiento discriminativo es la reducción del número de cálculos por dato a la entrada. En los ensayos se prueban ambos entrenamientos. El resultado es que no hay diferencias apreciables entre uno y otro: entrenar todos los filtros no acelera la convergencia, y no hacerlo no tiene efectos negativos sobre el comportamiento del filtro.

3.3. SIMULACIONES DEL ALGORITMO EN ESCALERA

3.3.1. Descripción de las simulaciones

Si el elemento que provoca la distorsión no lineal está situado a la entrada del canal (no linealidad sin memoria), el tipo de algoritmo lineal a aplicar será el filtro FIR con umbrales adaptativos, ya que todas las clases presentan una

distribución paralela: los diferentes hiperplanos óptimos[15] de discriminación se diferencian entre ellos únicamente en el término independiente: los hiperplanos son paralelos.

Sin embargo, esta solución no funcionará correctamente si el elemento no lineal está situado a la salida del canal (no linealidad con memoria). Incluso para pequeñas distorsiones no lineales este efecto puede observarse en las simulaciones que vienen a continuación.

Los parámetros de las simulaciones son:

- ruido gaussiano blanco aditivo con una SNR de 25 y 30 dB;
- señal PAM de 8 niveles $\{\pm1, \pm3, \pm5, \pm7\}$y símbolos uniformemente distribuidos;
- canal lineal de función de transferencia $H(z)=1+0.2z^{-1}$.
- No linealidad:

$$G(x) = \frac{1}{\tanh\left(\frac{1}{\xi}\right)} \tanh\left(\frac{x}{\xi}\right)$$

que simula el comportamiento de un dispositivo de potencia cuando la señal de salida tiene sólo componente en fase (véase Capítulo 1). Se escoge $\xi=10$, lo que significa que el canal se satura en esta amplitud.

Tal compresión degrada ostensiblemente las prestaciones del filtro FIR con umbrales equidistantes, pero además, si la no linealidad es con memoria, la degradación en las prestaciones del algoritmo FIR con umbrales adaptativos también es importante, como se observa en las simulaciones del Capítulo 2.

Se ensayan por separado las no linealidades con memoria y sin memoria. La no linealidad modelo es una saturación del tipo tangente hiperbólica con valor máximo 10.

Las funciones de coste comparadas son:

- Norma L_2 o error cuadrático con filtros de salida lineal;
- Norma L_2 o error cuadrático con filtros de salida tangente hiperbólica
- Norma entrópica

Se llevan a cabo tres medidas:

- Tasa de error en recepción. Para obtenerla se entrena el algoritmo en fase

[15] Para la familia de hiperplanos de discriminación de orden M en el espacio de datos, el óptimo es aquél que produce el mínimo error medio (para alguna medida de este error).

guiada durante 10.000 muestras. Se escoge este número porque garantiza que el algoritmo llegue al estado estacionario en todos los casos en que éste converge. Posteriormente se conmuta a funcionamiento dirigido por decisión y se prosigue con el algoritmo hasta que se producen 100 errores. Con ello se mide la tasa de error en la fase dirigida por decisión.

- Tiempo de convergencia. Se promedia el error cuadrático para 100 realizaciones independientes (véase Figura 3-6). Cada realización contiene 10.000 muestras. Sobre la gráfica del error promediado se mide el instante en que el valor de aquél ha recorrido el 90% de la distancia entre el error inicial y el error final. Consideramos este tiempo como el tiempo medio de convergencia del algoritmo.
- Capacidad de seguimiento de la variación del canal. Se mide cambiando la respuesta al impulso del canal $h[n]=\delta[n]+h_0\delta[n-1]$. Primero se entrena el algoritmo guiado durante 10.000 muestras y con $h_0=0,2$. Se conmuta a entrenamiento dirigido por decisión y, cuando han pasado 2.000 muestras más, se cambia el canal a $h_0=0.05$. Se cuenta el número de veces que el algoritmo falla. Se entiende que el algoritmo ha fallado cuando el número de errores después del cambio de canal es superior al 10%. Esto es válido porque esta tasa de fallos sólo se observa en condiciones de seguimiento para pasos de adaptación superiores a 0.05 y un canal con interferencia intersimbólica cuatro veces superior a la del canal que estamos utilizando en esta medida.

Las comparaciones se llevan a cabo para diferentes pasos de adaptación

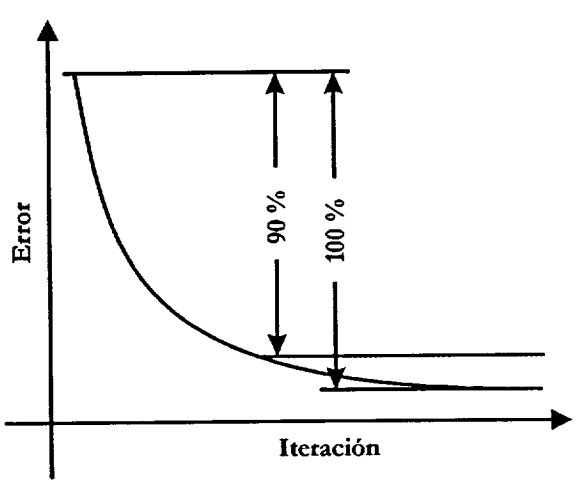


Figura 3-6. El tiempo de convergencia se mide como el tiempo necesario para que el error recorra el 90% del intervalo entre el error inicial y el error final. Para hacer la medida se promedian 100 realizaciones de 10000 iteraciones (muestras) cada una de ellas.

μ. El intervalo utilizado para μ va desde 10^{-4} hasta 10^{-1} en 12 intervalos equiespaciados en la escala logarítmica, que es la que utilizaremos en las gráficas.

El entrenamiento es por gradiente utilizando las tres funciones de coste descritas anteriormente. Se comparan sus prestaciones con las del filtro FIR con umbrales adaptativos.

Todas las variantes del algoritmo se prueban con los mismos conjuntos de datos y ruido: las simulaciones son idénticas

3.3.2. .Resultados de los algoritmos no discriminativos

3.3.2.a No linealidad sin memoria

La Figura 3-7 corresponde a la probabilidad de error del algoritmo en escalera para las tres funciones de coste vistas en el apartado 3.2.2.

Se observa que las prestaciones se degradan rápidamente a partir de $4 \cdot 10^{-4}$ y $6 \cdot 10^{-4}$ para SNR = 25 y 30 dB respectivamente. El sistema es más sensible a μ cuando la función de coste es entrópica o cuadrática con activación tangente

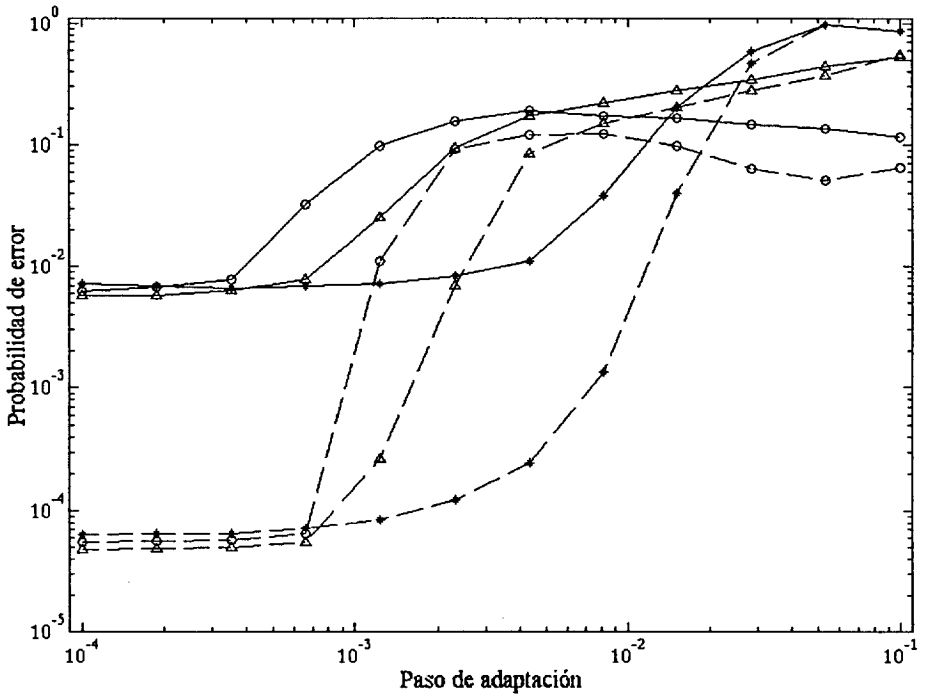


Figura 3-7. Probabilidad de error respecto del paso de adaptación para el algoritmo en escalera y para las tres funciones de coste: cuadrática con activación lineal (*), cuadrática con activación tangente hiperbólica (Δ) y entrópica (o). SNR=25 dB para las gráficas en línea continua y 30 dB para las gráficas en línea discontinua. No linealidad a la entrada del canal.

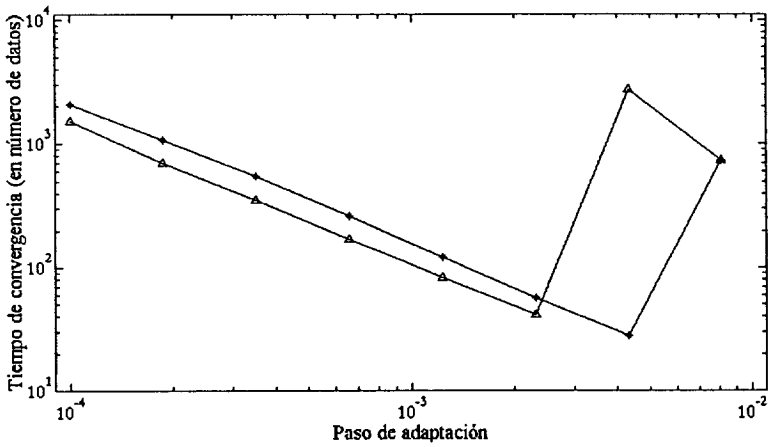


Figura 3-8. Tiempo de convergencia para el algoritmo en escalera versión no discriminativa y no linealidad sin memoria. Funciones de coste cuadrática lineal (*) y cuadrática con sigmoide (△). No se representa el tiempo de convergencia del coste entrópico porque es mayor que 10^4 muestras.

hiperbólica. El aumento de la tasa de fallos se debe en realidad a que para pasos de adaptación mayores a éstos el sistema diverge en algunas realizaciones, más cuanto mayor es el paso de adaptación. Por otro lado, y dentro del margen de funcionamiento aceptable se observa una tasa de fallos ligeramente menor en las funciones de coste que utilizan activación tangente hiperbólica.

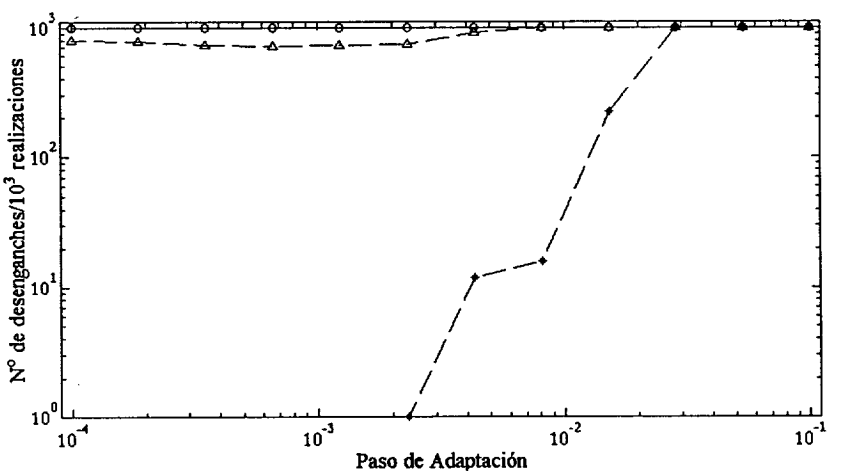


Figura 3-9. Capacidad de seguimiento de canal no estacionario medida como el número de desenganches por 1000 realizaciones. La respuesta al impulso del canal es $h[n]=\delta[n]+h_0\delta[n-1]$. Se entrena el algoritmo en modo guiado durante 10.000 muestras y con $h_0 = 0{,}2$; luego se conmuta a entrenamiento dirigido por decisión y se deja durante 2.000 muestras más. En ese instante se conmuta a $h_0 = 0{,}05$. Se repite para 1000 realizaciones, contando el número de desenganches. Coste cuadrático con activación lineal (*), cuadrático con activación tangente hiperbólica (△) y entrópico (o).

Compárense estos resultados con los del filtro FIR de umbrales adaptativos del Capítulo 2. La primera conclusión que debemos extraer es que el algoritmo en escalera es inestable para la mayoría de pasos de adaptación para los que el filtro FIR de coeficientes adaptativos funciona correctamente. Sin embargo, la ventaja principal sobre el filtro FIR con umbrales adaptativos es que las prestaciones en canal con no linealidad con memoria no se degradan, como veremos en los siguientes apartads.

En la Sección 1.1 se presentará un método para hacer que la convergencia del algoritmo en escalera sea igual a la del filtro FIR con umbrales adaptativos.

La Figura 3-8 muestra el tiempo de convergencia del algoritmo medido como el número medio de muestras necesarias para que el error llegue a un valor 10 veces superior al error final. No se representa el tiempo de convergencia para la función de coste entrópica, dado que es mayor que 10^4 muestras. Se observa un tiempo de convergencia entre 100 y 1000 muestras para el intervalo del paso de adaptación que produce una tasa de error aceptable.

Como se observa en la Figura 3-9, la capacidad de seguimiento del sistema es malo, excepto para el coste cuadrático con activación lineal. Para los costes cuadrático con tangente hiperbólica y entrópico, las prestaciones son malas: el desenganche se produce casi siempre. El coste cuadrático, en cambio, presentó cero desenganches sobre las 1000 realizaciones para pasos de adaptación entre 10^{-4} y $4 \cdot 10^{-3}$.

3.3.2.b No linealidad con memoria

En las simulaciones cuyos resultados se presentan aquí se mantienen las mismas condiciones que en el anterior experimento, excepto porque la no linealidad se sitúa tras la parte lineal del canal: se tiene ahora una no linealidad con memoria.

Aunque las prestaciones del algoritmo en escalera en su versión inicial no son aceptables, se ensaya este algoritmo para un canal con no linealidad con memoria, a fin de comprobar si efectivamente tiene ventaja con respecto al algoritmo FIR con umbrales adaptativos; lo que ocurre, como se observa en la Figura 3-10: utilizando este algoritmo, y con un paso de adaptación en el intervalo desde 10^{-4} hasta $3 \cdot 10^{-4}$ la tasa de error está muy por debajo de la tasa de error del algoritmo FIR con umbrales adaptativos.

La mejora es evidente. Sin embargo, utilizando pasos de adaptación mayores, las prestaciones del algoritmo en escalera se degradan gravemente en tanto que las del FIR se mantienen razonablemente.

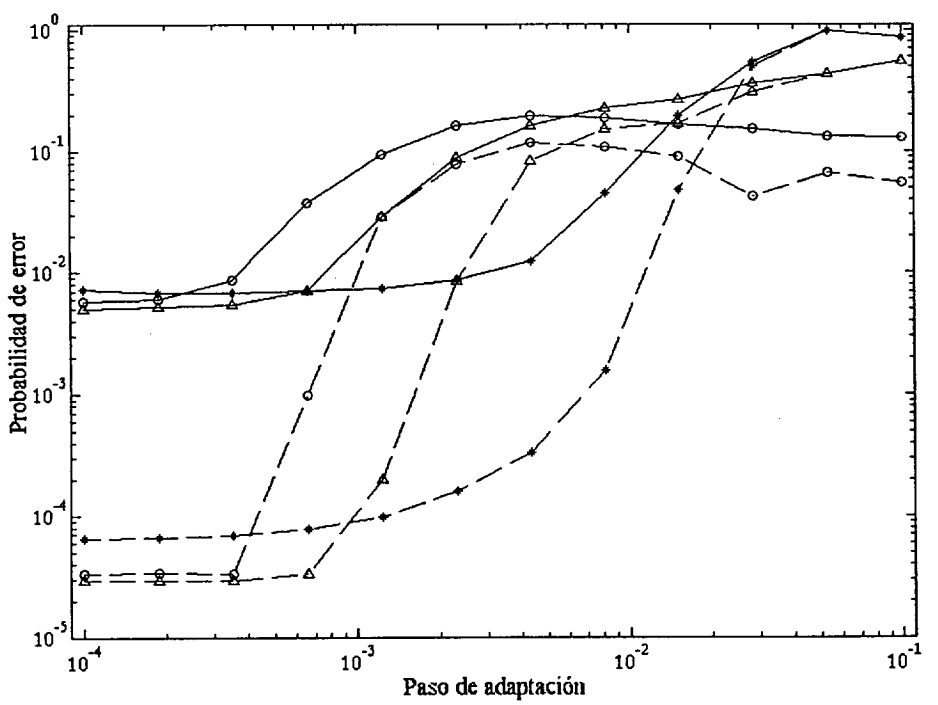


Figura 3-10. Probabilidad de error respecto del paso de adaptación para el algoritmo en escalera y para las tres funciones de coste cuadrática con activación lineal (*), cuadrática con activación tangente hiperbólica (△) y entrópica (o). SNR=25 dB para las gráficas en línea continua y 30 dB para las gráficas en línea discontinua. No linealidad con memoria.

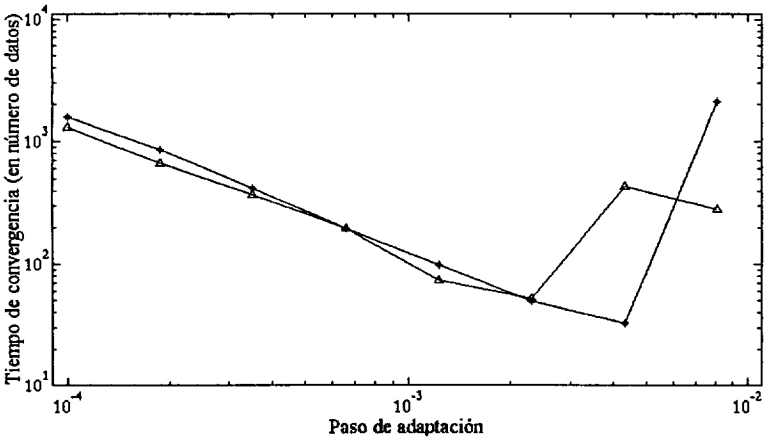


Figura 3-11. Tiempo de convergencia para la versión no discriminativa y no linealidad con memoria.. Coste cuadrático con activación lineal (*) y cuadrático con activación tangente hiperbólica (△). El coste entrópico tiene un tiempo de convergencia fuera de los márgenes de la gráfica.

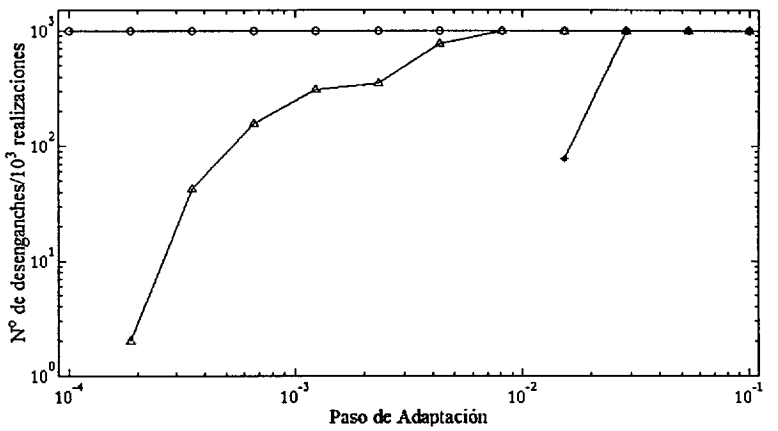


Figura 3-12. Capacidad de seguimiento de canal no estacionario medida como el número de desenganches por 1000 realizaciones. Coste cuadrático con activación lineal (*), cuadrático con activación tangente hiperbólica (△) y entrópica (o).

Véanse los tiempos de convergencia en la Figura 3-11 para diferentes pasos de adaptación. Son similares a los del filtro FIR con umbrales adaptativos, como es de esperar. Los datos correspondientes al coste entrópico están fuera de los límites de la gráfica.

La capacidad de seguimiento del coste entrópico sigue siendo inaceptable (Figura 3-12), aunque para las otras dos funciones de coste y paso de adaptación de 10^{-4} el número de desenganches fue cero en las 1000 realizaciones. En realidad, para el coste cuadrático se computan cero desenganches para el intervalo 10^{-4} a 10^{-2}, que puede decirse que será el intervalo de convergencia aceptable.

3.3.3. Versión discriminativa

3.3.3.a No linealidad sin memoria

Para la versión discriminativa el algoritmo con coste entrópico no converge. Las otras dos funciones de coste no se ven afectadas por la reducción de muestras de entrenamiento (Figura 3-13) en cuanto a que hay convergencia en el mismo intervalo que antes. El número de desenganches es casi exactamente el mismo que en la versión no discriminativa (Figura 3-14). En cuanto a la velocidad de convergencia, ésta aumenta para el coste cuadrático con las dos activaciones: por ejemplo, para $\mu=10^{-4}$ y activación lineal, el número de muestras para llegar a la convergencia se multiplica por 2 y casi se triplica con $\mu=10^{-3}$ (Figura 3-15).

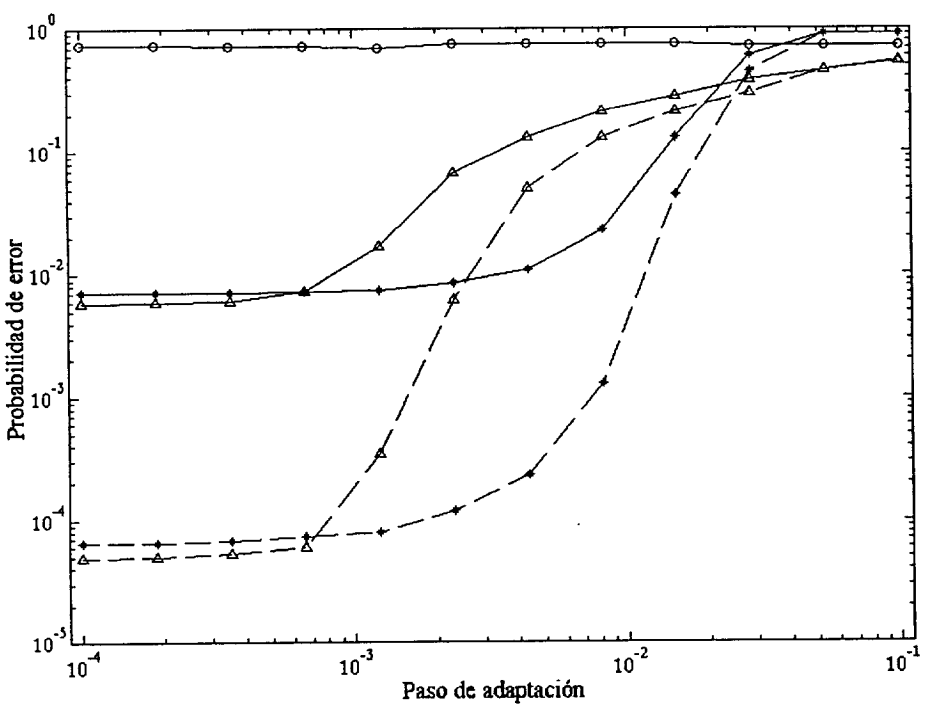


Figura 3-13. Tasa de error para el algoritmo en escalera, versión discriminativa. No linealidad sin menoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o). Las gráficas en trazo continuo corresponden a una SNR de 25 dB y las gráficas en trazo discontinuo a 30 dB.

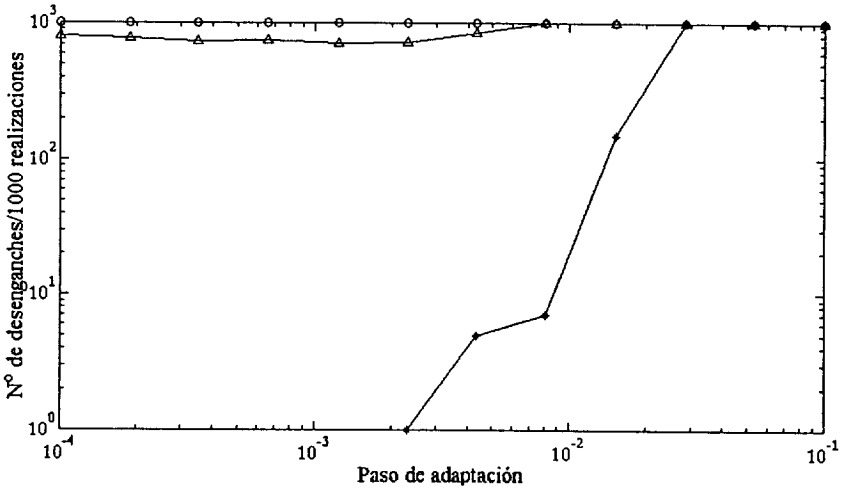


Figura 3-14. Número de desenganches en 1000 iteraciones para el algoritmo en escalera discriminativo. No linealidad sin memoria. Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o).

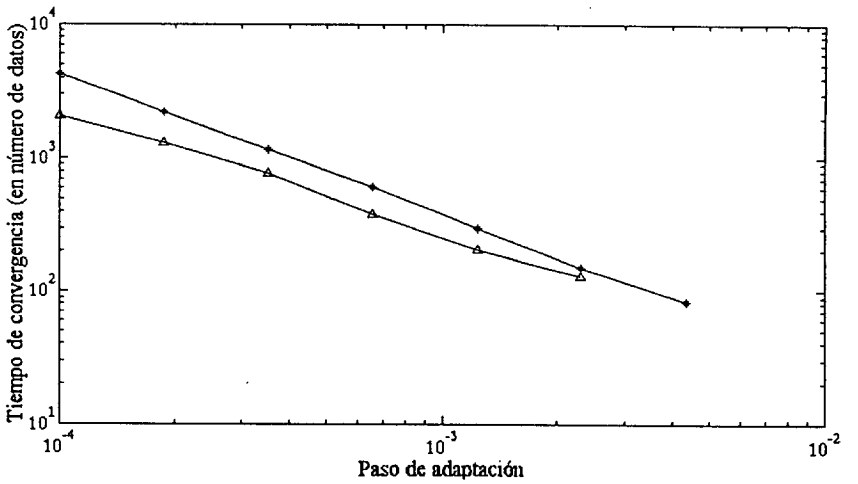


Figura 3-15. Tiempo de convergencia para el algoritmo en escalera versión discriminativa. No linealidad sin memoria. Coste cuadrático con activación lineal (*) y cuadrático con activación tangente hiperbólica (△).

Al igual que en la versión no discriminativa, el algoritmo con coste entrópico no puede seguir las variaciones del canal en ninguna de las 1000 realizaciones.

3.3.3.b No linealidad con memoria

Se repiten los experimentos del apartado 3.3.2.b pero para la versión discriminativa del algoritmo.

Las prestaciones en cuanto a tasa de error (Figura 3-16) son las mismas que para las versiones no discriminativas para coste cuadrático con y sin activación tangente hiperbólica. El coste entrópico no converge, igual que con no linealidad sin memoria.

Obsérvese, sin embargo, la diferencia entre la tasa de error mínima de las dos versiones del coste cuadrático con activación lineal y activación tangente hiperbólica. El segundo consigue menos de la mitad de errores que el primero. Esta mejora se debe a que la activación tangente hiperbólica hace que los datos que están muy lejos de la frontera no tomen parte en el entrenamiento de la frontera, al resultar su error muy pequeño. La frontera se adapta en función de los datos que están más cerca, que en este caso son sólo los de una zona muy pequeña del espacio (obsérvese la Figura 1-8, pág 18 para una visión intuitiva del espacio de datos en este canal).

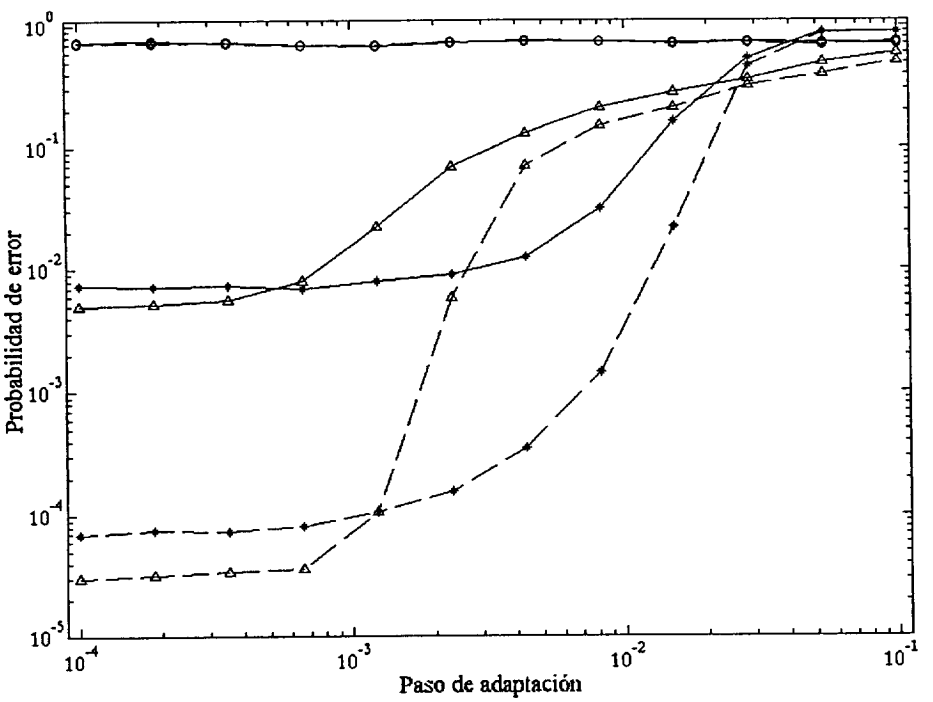


Figura 3-16. Tasa de error para el algoritmo en escalera, versión discriminativa. No linealidad con memoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (Δ) y coste entrópico (o). Las gráficas en trazo continuo corresponden a una SNR de 25 dB y las gráficas en trazo discontinuo a 30 dB.

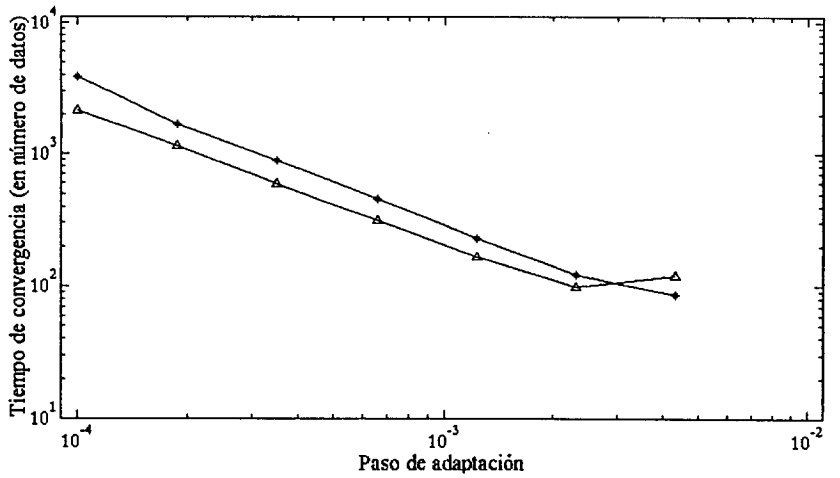


Figura 3-17. Tiempo de convergencia del algoritmo, versión discriminativa. No linealidad con memoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o).

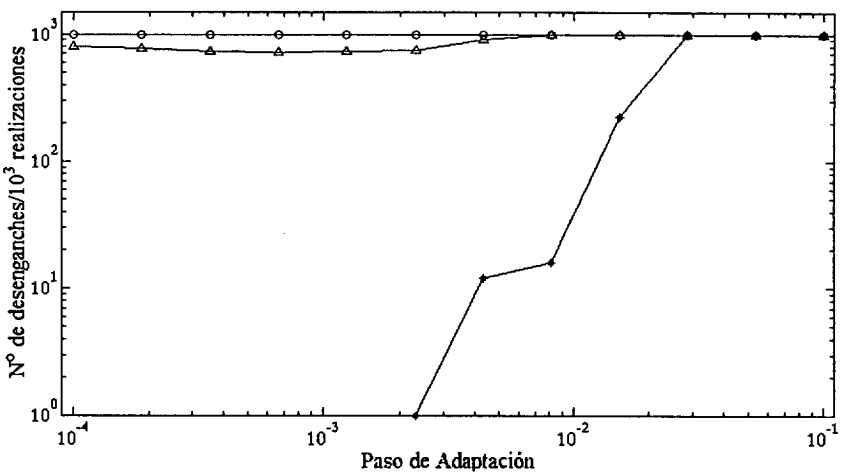


Figura 3-18. Número de desenganches en 1000 iteraciones para el algoritmo en escalera discriminativo. No linealidad con memoria. Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o).

En cuanto a la velocidad de convergencia, aumenta entre el doble y el triple (Figura 3-17). La tasa de desenganches frente al canal no estacionario es muy mala con coste cuadrático y activación tangente hiperbólica. Las prestaciones son similares a las del algoritmo en canal con no linealidad sin memoria.

3.4. DISCUSIÓN

La tasa de error es mucho más sensible al paso de adaptación para el algoritmo en escalera que para el filtro FIR. El algoritmo se vuelve inestable para pasos de adaptación superiores a $4 \cdot 10^{-4}$ para SNR=25 dB y $6 \cdot 10^{-4}$ para SNR=30 dB. Además, para la versión discriminativa, el coste entrópico no converge.

En lo que respecta a las ventajas del algoritmo en escalera, se ve claramente que para no linealidad con memoria el filtro FIR deja de funcionar correctamente, mientras que el algoritmo en escalera es capaz de igualar, siempre que el paso de adaptación sea pequeño. Es necesario, por tanto, mejorar el algoritmo para que acepte pasos de adaptación mayores, lo que producirá mayores velocidades de convergencia.

El algoritmo en escalera es mucho menos robusto que el filtro FIR en cuanto el canal es no estacionario. El filtro FIR presentó cero desenganches en las 1000 simulaciones que se llevaron a cabo para las tres funciones de coste. Sin embargo, el algoritmo en escalera también es capaz de seguir las variaciones del canal con un 0 % de desenganches si la función de coste es cuadrática con activación lineal. Por ello se deduce una alternativa: utilizar una combinación de costes que pondere uno u otro en función de la naturaleza estacionaria o no del canal. Para detectar la no estacionaridad hay que estimar el error cuadrático a la salida del filtro del clasificador central, que es el único que clasifica todas las muestras. Si este error es pequeño, se usará el coste entrópico o una combinación que pondere más el coste entrópico cuanto menor sea el error. En otro caso, debe utilizarse el coste cuadrático o ponderar la combinación a su favor.

Cuando el ruido de observación sea alto, el error también lo será, y no se conmutará nunca al entrópico, pero en condiciones de SNR baja (véanse las anteriores gráficas para SNR=25 dB) todos los algoritmos presentan iguales prestaciones en tasa de error.

La velocidad de convergencia aumenta con el uso del algoritmo en escalera en los casos en que sus prestaciones son comparables con las del filtro FIR (es decir, cuando la no linealidad es sin memoria). Además, el uso de activación tangente hiperbólica parece proporcionar alguna ventaja sobre la activación lineal.

Las tablas que se adjuntan a continuación son un resumen de los

resultados obtenidos para el algoritmo en escalera con SNR=30 dB y su comparación con las prestaciones del filtro FIR.

Tabla 3-1. Comparación de las tasas de error del algoritmo en escalera y el FIR con umbrales adaptativos para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN).; no linealidad sin memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	$6{,}6\cdot10^{-5}$	$5{,}0\cdot10^{-5}$	$5{,}6\cdot10^{-5}$	$8{,}0\cdot10^{-5}$	$1{,}5\cdot10^{-4}$	$1{,}9\cdot10^{-3}$
Escalera, Disc.	$6{,}6\cdot10^{-5}$	$5{,}1\cdot10^{-5}$	0,75	$7{,}0\cdot10^{-5}$	$1{,}5\cdot10^{-4}$	0,75
FIR	$6{,}5\cdot10^{-5}$	$5{,}0\cdot10^{-5}$	$5{,}6\cdot10^{-5}$	$4{,}0\cdot10^{-5}$	$4{,}0\cdot10^{-5}$	$7{,}0\cdot10^{-5}$

Tabla 3-2. Comparación de las tasas de error del algoritmo en escalera y el FIR con umbrales adaptativos para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN);no linealidad con memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	$6{,}7\cdot10^{-5}$	$3{,}0\cdot10^{-5}$	$3{,}4\cdot10^{-5}$	$9{,}1\cdot10^{-5}$	$1{,}0\cdot10^{-4}$	$8{,}8\cdot10^{-3}$
Escalera, Disc.	$7{,}5\cdot10^{-5}$	$3{,}2\cdot10^{-5}$	0,75	$9{,}8\cdot10^{-5}$	$7{,}4\cdot10^{-5}$	0,75
FIR	$2{,}5\cdot10^{-4}$	$2{,}1\cdot10^{-3}$	$3{,}0\cdot10^{-3}$	$2{,}8\cdot10^{-4}$	$1{,}8\cdot10^{-3}$	$3{,}0\cdot10^{-3}$

Tabla 3-3. Comparación de los tiempos de convergencia en número de muestras del algoritmo en escalera y el FIR con umbrales adaptativos para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN); no linealidad sin memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	2160	1650	$>10^4$	620	480	-
Escalera, Disc.	2250	1630	-	310	225	-
FIR	1800	4500	$>10^4$	350	1100	3800

Tabla 3-4. Comparación de las velocidades de convergencia del algoritmo en escalera y el FIR con umbrales adaptativos para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN); no linealidad con memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	1850	1580	$>10^4$	540	470	-
Escalera, Disc.	2950	2270	-	940	750	-
FIR	7500	-	-	600	-	-

Tabla 3-5. Comparación de los porcentajes de desenganche frente a canal no estacionario del algoritmo en escalera y el FIR con umbrales adaptativos para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN); no linealidad sin memoria.

Paso de adaptación	$2 \cdot 10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	0 %	76 %	100 %	0 %	79 %	10 %
Escalera, Disc.	0 %	79 %	100 %	0 %	79 %	100 %
FIR	0 %	0 %	0 %	0 %	0 %	0 %

Tabla 3-6. Comparación de los porcentajes de desenganche frente a canal no estacionario del algoritmo en escalera y el FIR con umbrales adaptativos para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN); no linealidad con memoria.

Paso de adaptación	$2 \cdot 10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	0 %	2,6 %	100 %	0 %	24 %	100 %
Escalera, Disc.	0 %	1,4 %	100 %	0 %	37 %	100 %
FIR	0 %	0 %	0 %	0 %	0,5 %	0,5 %

En las tablas 3-1 y 3-2 se observa que el coste entrópico da la menor tasa de errores si el paso de adaptación es suficientemente bajo, pero si aumenta el paso de adaptación o se utiliza entrenamiento discriminativo, este coste no converge. El coste cuadrático funciona algo mejor, y además converge con entrenamiento discriminativo, pero es igualmente sensible al paso de adaptación.

Las tablas 3-3 y 3-4 muestran los tiempos de convergencia. El entrenamiento discriminativo no disminuye significativamente la velocidad de convergencia. El coste entrópico, sin embargo, la tiene muy alta.

Por último, se ven, en las tablas 3-5 y 3-6, los porcentajes de desenganche en un canal no estacionario. Vemos que el coste cuadrático con activación lineal es muy robusto. Con activación tangente hiperbólica la robustez es menor, aunque si se baja el paso de adaptación, la tasa de desenganches disminuye. El coste entrópico es inestable. Vemos que, sin embargo, en le filtro FIR, la tasa de desenganches con coste entrópico es razonablemente baja.

La función de coste entrópica toma las salidas como una aproximación a las probabilidades a posteriori. Esta función no converge adecuadamente porque las condiciones iniciales de funcionamiento (la respuesta al impulso inicial de los filtros es $h[n] = \delta[n]$) hacen que la aproximación no sea buena en los filtros más exteriores, con lo que la medida de la entropía no es válida en ellos. No obstante, el filtro central está mejor situado (el peso independiente del filtro es nulo al

inicio, y, por simetría, en la convergencia lo sigue siendo; los datos de diferentes clases están más separados unos de otros, con lo que la tasa inicial de errores es menor) y su medida de la entropía es más precisa: este filtro sí tiende a converger.

El uso de distintos pasos de adaptación para los diferentes filtros que intervienen en el algoritmo en escalera producirá un aumento de la estabilidad y una mejora en la convergencia, en lo que respecta a tasa de errores. Pero si se quiere aumentar la velocidad de convergencia, no es suficiente el uso de pasos de adaptación diferentes. Es necesario utilizar la convergencia del filtro central para hacer que los demás filtros converjan.

El capítulo siguiente trata de la forma en que se debe modificar el algoritmo en escalera para que su comportamiento sea mejor con pasos de adaptación mayores y para que sea estable cuando se aplica el coste entrópico.

4. ALGORITMO EN ESCALERA MODIFICADO

4.1. MEJORA DE LA ESTABILIDAD DEL ALGORITMO EN ESCALERA

El algoritmo en escalera converge sólo para pasos de adapación pequeños y además, el coste entrópico no converge en muchos casos, siendo poco estable frente a canales no estacionarios.

Con la modificación que se describe a continuación el sistema converge para pasos de adaptación diez veces mayores. Además, la forma en que se lleva a cabo el entrenamiento que se presenta en esta sección tiene como característica un mayor aprovechamiento de los datos, lo que resulta en una mayor velocidad de adaptación, que trae consigo robustez del sistema frente a canales no estacionarios.

El entrenamiento del sistema es el mismo para cualquiera de los canales no lineales que se describen en este trabajo; sin embargo, se tratan por separado primero el canal lineal, después el canal con no linealidad a la entrada (no linealidad sin memoria), y por último el canal con no linealidad a la salida (no linealidad con memoria).

En este capítulo se resume el análisis llevado a cabo a fin de justificar en forma analítica la modificación que se propone; el desarrollo detallado del análisis se ha recogido en el Anexo 4.2.

4.1.1. Notación

El vector de datos de entrada al igualador en el instante n tiene la forma

$$\mathbf{r}[n] = \{r[n], \cdots, r[n-M+1], 1\}^T$$

El término constante se corresponde con el término independiente del vector de pesos.

El algoritmo en escalera distribuye los datos de entrada hacia los diferentes clasificadores. Llamamos $\mathbf{r}^{b_0 b_1 \cdots b_i}$ al conjunto de vectores de entrada al clasificador correspondiente de la capa i del igualador.

Por otro lado, $\mathbf{w}^{b_0 b_1 \cdots b_i}$ es el vector de pesos del filtro cuya entrada es el conjunto anterior

Respecto al ruido, supondremos en todos los casos que los datos de entrada a cualquiera de los filtros están afectados por ruido gaussiano blanco aditivo de media nula, y también que se mantienen siempre condiciones de bajo ruido.

4.1.2. Solución óptima para canal con no linealidad sin memoria

Llamamos solución óptima de cada filtro a aquélla que minimiza el error cuadrático medio[16] de la salida del filtro. Ya que cada filtro es entrenado por un conjunto de datos disjunto de los otros conjuntos de datos de la misma capa del igualador, la solución óptima depende sólo de éstos. Para un filtro cuyo vector de coeficientes es $\mathbf{w}^{b_0 b_1 \cdots b_i}$ y cuya salida clasifica los datos $\mathbf{r}^{b_0 b_1 \cdots b_i}$, la solución que minimiza el error cuadrático medio es la que verifica $\mathbf{R}^{b_0 b_1 \cdots b_i} \mathbf{w}^{*b_0 b_1 \cdots b_i} = \mathbf{p}^{b_0 b_1 \cdots b_i}$, donde $\mathbf{R}^{b_0 b_1 \cdots b_i}$ es la autocorrelación de los datos y $\mathbf{p}^{i,j}$ es el vector de correlación cruzada entre los datos y la salida deseada. Si se calculan $\mathbf{R}^{b_0 \cdots b_j}$ y $\mathbf{p}^{b_0 \cdots b_j}$ se observan las relaciones entre los coeficientes del vector óptimo $\mathbf{w}^{*b_0 b_1 \cdots b_i}$ para cada uno de los filtros.

El desarrollo de la matriz de autocorrelación para datos procedentes de un canal lineal está en el apartado A4.2.3 del Anexo 4.2. Si se sustituye en la ecuación del filtro óptimo se tiene

---

[16] El coste error cuadrático con activación lineal es sólo uno de los tres que se utilizan en esta tesis, y el que produce peores prestaciones en cuanto a probabilidad de error (si el paso de adaptación es superior a 10^{-3} el incremento en probabilidad de error respecto de los otros costes es significativo). Deberíamos diferenciar tres tipos de solución óptima, cada una de las cuales minimizando uno de los tres costes. No obstante, el cálculo de la solución óptima sólo es sencilla para el coste cuadrático con activación lineal, donde se ven involucradas funciones lineales y cuadráticas. Para las otras dos funciones de coste se supondrá que la solución óptima es parecida a la solución que minimiza el error cuadrático medio.

$$\begin{bmatrix} \|\mathbf{h}\|^2\sigma_x^2 + h_0^2\left(m_X^{2\,b_0\ldots b_j} + \sigma_X^{2\,b_0\ldots b_j} - \sigma_x^2\right) & \rho_h^2[0,1]\sigma_x^2 & \cdots & \rho_h^2[0,M-1]\sigma_x^2 & h_0 m_X^{b_0\ldots b_j} \\ \rho_h^2[1,0]\sigma_x^2 & \ddots & & & \\ \vdots & & \ddots & & \vdots \\ \rho_h^2[M-1,0]\sigma_x^2 & & & \ddots & 0 \\ h_0 m_X^{b_0\ldots b_j} & 0 & \cdots & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} w_0^{b_0\ldots b_j} \\ \vdots \\ \vdots \\ \vdots \\ w_M^{b_0\ldots b_j} \end{bmatrix} = \begin{bmatrix} p \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

siendo:

$$\sigma_x^2 = E\left\{x^2[n]\right\}$$

$$m_X^{b_0\ldots b_j} = E\left\{\left(x^{b_0\ldots b_j}\right)\right\}$$

$$\sigma_X^{2\,b_0\ldots b_j} = E\left\{\left(x^{b_0\ldots b_j} - m_X^{b_0\ldots b_j}\right)^2\right\}$$

$$\rho_h^2[p,q] = \sum_{j=p-q}^{N-1} h_{q-p+j} h_j^* \qquad p > q$$

$$p = E\left\{d \cdot r^{i,j}[n]\right\}$$

Nótese que $E\left\{d \cdot r^{i,j}[n-k]\right\} = 0$ si $k > 0$, ya que sólo el primer término del vector de datos contiene el símbolo a detectar.

Para el último coeficiente del vector (término independiente) la solución es $w_M^{i,j^*} = -w_0^{i,j^*} m_{x^{i,j}}$. Sustituyendo esta solución en la ecuación matricial se obtiene un sistema de M-1 ecuaciones y M-1 incógnitas. Al ser nulos todos los elementos del vector a la derecha de igualdad, las ecuaciones de las filas 2 a M-1 definen una línea recta que pasa por el origen.

Estas ecuaciones son iguales para las matrices de autocorrelación de todos los filtros: por lo tanto, el vector director de la recta definida por estas ecuaciones es el mismo para todos los filtros. Si se despeja la ecuación de la fila M, la intersección de la recta con el plano de la ecuación 1 proporciona la solución a los coeficientes $w_0^{b_0\ldots b_j} \cdots w_{M-1}^{b_0\ldots b_j}$. El valor del coeficiente M sale despejando la última ecuación y vale $w_M^{b_0\ldots b_j} = w_0^{b_0\ldots b_j} m_X^{b_0\ldots b_j}$

Cada una de las matrices de autocorrelación correspondientes a los filtros del igualador en escalera tiene un valor diferente en la primera fila, lo que

significa que la solución óptima será diferente para cada filtro. Pero como la recta definida por las ecuaciones 2 a M-1 pasa por el origen y tioene el mismo vector director cualquiera que sea la matriz de autocorrelación, todas las soluciones son linealmente dependientes.

Sea

$$\hat{\mathbf{w}}^{*b_0 \cdots b_j} = \frac{\mathbf{w}^{*b_0 \cdots b_j}}{w^{*b_0 \cdots b_j}_0} = \left[1 \quad \frac{w^{*b_0 \cdots b_j}_1}{w^{*b_0 \cdots b_j}_0} \quad \frac{w^{*b_0 \cdots b_j}_{M-1}}{w^{*b_0 \cdots b_j}_0} \quad \cdots \quad h_0 m_{x^{b_0 \cdots b_j}} \right]^T$$

el vector correspondiente al filtro óptimo para el conjunto de datos $x^{b_0 \cdots b_j}$. Como los coeficientes $w_0^{b_0 \cdots b_j} \cdots w_{M-1}^{b_0 \cdots b_j}$ son proporcionales, se puede reescribir el vector $\hat{w}^{*b_0 \cdots b_j}$ en función del vector correspondiente al filtro central $\mathbf{w}^*$:

$$\hat{\mathbf{w}}^{*b_0 \cdots b_j} = \frac{\mathbf{w}^{*b_0 \cdots b_j}}{w^{*b_0 \cdots b_j}_0} = \left[1 \quad \frac{w^*_1}{w^*_0} \quad \frac{w^*_{M-1}}{w^*_0} \quad \cdots \quad h_0 m_{x^{b_0 \cdots b_j}} \right]^T \qquad (4\text{-}1)$$

Para el canal no lineal sin memoria, el anterior resultado es válido, con la salvedad de que hay que sustituir $m_{x^{i,j}}$ por $m_{F(x^{i,j})} = E\{F(x^{i,j})$ siendo F(·) la parte no lineal del canal:

$$\hat{\mathbf{w}}^{*b_0 \cdots b_j} = \frac{\mathbf{w}^{*b_0 \cdots b_j}}{w^{*b_0 \cdots b_j}_0} = \left[1 \quad \frac{w^*_1}{w^*_0} \quad \frac{w^*_{M-1}}{w^*_0} \quad \cdots \quad h_0 m_{F(x^{b_0 \cdots b_j})} \right]^T \qquad (4\text{-}2)$$

Una modificación del algoritmo consiste en entrenar sólo el vector correspondiente al filtro central **w**, imponer los coeficientes 1 hasta M-1 de ese filtro en todos los demás, y calcular mediante gradiente sólo el término $w_M^{b_0 \cdots b_j} = w_0^{b_0 \cdots b_j} m_{x^{b_0 \cdots b_j}}$ de cada filtro. Como la información acerca de la clase a la que pertenece el dato está contenida en el signo de la salida, poco importa que ésta esté escalada, y por tanto la solución es válida. Esto, obviamente, equivale al filtro FIR con umbrales adaptativos, que, como se ha visto en el Capítulo 2, produce buenos resultados con canal no lineal sin memoria.

4.1.3. Solución óptima para canal con no linealidad con memoria

Si la no linealidad está situada a la salida del canal (no linealidad con memoria) los filtros no son paralelos: se ha comprobado experimentalmente que

la solución propuesta en el apartado anterior no es válida para el caso 8-PAM (véase para ello el Capítulo 2).

Cuando el canal es no lineal con memoria las matrices de autocorrelación de los datos a la entrada de cada uno de los filtros ya no tiene la misma estructura que en un canal lineal o no lineal sin memoria, por lo que, a primera vista, ya no es posible propagar el entrenamiento del filtro central a los demás. Pero un cálculo detallado de las matrices de autocorrelación da como resultado que esto no es totalmente cierto. Este cálculo está en el anexo 4.2, apartado A4.2.5; aquí se expone el resultado.

La matriz de autocorrelación de la entrada a cada uno de los filtros puede escribirse en la forma

$$\mathbf{R}^{b_0 \cdots b_i} = \sum_{l=1}^{\infty} \sum_{m=1}^{\infty} G^{(2l-1)} G^{(2m-1)} \mathbf{R}_{2l-1,2m-1}^{b_0 \cdots b_i} \tag{4-3}$$

donde $G^{(k)}$ es la derivada de orden k de la no linealidad G(·) y

$$\mathbf{R}_{2l-1,2m-1}^{b_0 \cdots b_i} = \sum_{l=1,m=1}^{\infty} G^{(2l-1)} G^{(2m-1)} E\left(\begin{bmatrix} \left(\mathbf{x}^{b_0 \cdots b_i\,T}\mathbf{h}\right)^{2l-1} \\ \vdots \\ \left(\mathbf{x}^{T}[n-M+1]\mathbf{h}\right)^{2l-1} \\ \left(G^{(2l-1)}\right)^{-1} \end{bmatrix} \begin{bmatrix} \left(\mathbf{x}^{b_0 \cdots b_i\,T}\mathbf{h}\right)^{2m-1} \\ \vdots \\ \left(\mathbf{x}^{T}[n-M+1]\mathbf{h}\right)^{2m-1} \\ \left(G^{(2m-1)}\right)^{-1} \end{bmatrix}^{T} \right) \tag{4-4}$$

Dos conclusiones se derivan del desarrollo de esta matriz:

- las matrices de autocorrelación son diferentes para entrenamiento discriminativo y no discriminativo: producen soluciones diferentes.
- los planos de separación no son paralelos, aunque sus diferencias son pequeñas incluso para niveles de saturación importantes.

Se observa que las ecuaciones de las filas 2 a M-1 no son iguales para todos los filtros (y por lo tanto los planos definidos por las ecuaciones matriciales $\mathbf{R}^{b_0 b_1 \cdots b_i}\mathbf{w}^{*\,b_0 b_1 \cdots b_i} = \mathbf{p}^{b_0 b_1 \cdots b_i}$ ya no son espacios afines al planos del filtro central), aunque sólo difieren en el término de la primera columna. Además, todos los elementos de la matriz serán pequeños excepto los de la diagonal . Si pueden despreciarse los términos más allá de l,m=3 de la ecuación (4-3), puede verse que el ángulo entre la recta definida por la ecuación 2ª de la matriz de autocorrelación del filtro central y la del filtro más exterior está en torno a los 6° para entrenamiento discriminativo y 1,5° para entrenamiento no discriminativo Este cálculo está hecho para un "back-off" de 0 dB y una interferencia

intersimbólica del 20% de la amplitud. El resultado no depende del número de niveles del alfabeto. Los ángulos entre las otras rectas serán menores a éste. El parecido entre las diferentes soluciones permite introducir la siguiente variante del entrenamiento.

4.1.4. Entrenamiento general

La idea del entrenamiento se basa en que una primera aproximación es hacer que todos sean paralelos y con la misma orientación que el plano central situado según la solución de Wiener. Luego se entrenan los filtros contiguos, que ya tendrán una posición cercana a la correcta, pero con un paso de adaptación que asegure la estabilidad. La manera de medir si el plano central ha convergido a la solución de Wiener es ver que la media de las estimaciones del gradiente es cero, lo cual lleva a realizar el algoritmo de la siguiente manera: la actualización de cada uno de los planos es la suma ponderada del gradiente del coste medido en el filtro que se está actualizando más el gradiente del coste medido en el filtro central. La actualización del filtro i será:

$$\mathbf{w}_m^{b_0 \dots b_j}[n+1] = \mathbf{w}_m^{b_0 \dots b_j}[n] + \eta\mu\Delta w_m^{b_0 \dots b_j} + \left(1-\eta^{b_0 \dots b_j}\right)\mu\Delta w_m, \quad m < M$$

$$\mathbf{w}_M^{b_0 \dots b_j}[n+1] = \mathbf{w}_M^{b_0 \dots b_j}[n] + \mu\Delta w_M^{b_0 \dots b_j} \tag{4-5}$$

para los términos dependientes en la ecuación del filtro, siendo μ el paso de adaptación y $\eta^{b_0 \dots b_j}$ la proporción del entrenamiento del discriminante central que se aplica a cada uno de los filtros contiguos. Basta con asegurar que $\left(1-\eta^{b_0 \dots b_j}\right)\mu$ sea suficientemente pequeño para asegurar la estabilidad del entrenamiento de la recta i. Para el término independiente la actualización es independiente, puesto que este término no tiene información que esté contenida en el plano central al ser un término que depende del valor medio de las señales a discriminar, el cual será diferente para cada grupo de datos.

Cuando el gradiente del vector $\mathbf{w}$ (que es el del filtro central) haya convergido al mínimo coste, la componente del gradiente en la dirección de este peso será cero, momento a partir del cual la ecuación actualiza el vector $\mathbf{w}^{b_0 \dots b_j}$ de forma independiente (ajuste fino). Cuando el vector $\mathbf{w}$ converja, los siguientes se actualizarán de forma independiente a velocidad menor hasta llegar a la convergencia de todo el sistema. Con esto se consigue que el sistema se acerque rápidamente a la convergencia, ya que para la etapa en que los filtros se actualizan con el filtro central, se utilizan todos los datos en el entrenamiento. La convergencia "fina" es más lenta, al utilizarse sólo los datos locales a cada uno de

los planos de discriminación y un paso de adaptación igual a $\left(1-\eta^{b_0 \cdots b_i}\right)\mu$, mucho más pequeño que μ. Es por ello que debe establecerse una solución de compromiso entre velocidad de convergencia y estabilidad. Para adaptar el plano de discriminación $b_0...b_j$ el valor de $\eta^{b_0 \cdots b_i}$ debe escogerse menor para planos exteriores y mayor para planos próximos al central. Esto es debido a que los planos próximos serán muy parecidos. En los ensayos se ha probado el algoritmo con señal 8-PAM y canal $H(z)=1+h_0z^{-1}$. Los filtros serán, por tanto, cuatro de primer orden, y se utilizan los parámetros $\eta^{11}=\eta^{00}=10^{-3}$, $\eta^{0}=\eta^{1}=10^{-2}$, $\eta^{01}=\eta^{01}=10^{-1}$ y $10^{-4}<\mu<10^{-2}$. Se observa cómo, efectivamente, la estabilidad aumenta significativamente. También se hacen pruebas de estabilidad frente a canales no estacionarios, con buenos resultados frente a los de los sistemas ensayados sin esta mejora. También puede verse cómo la velocidad de convergencia aumenta para todas las funciones de coste ensayadas. Sin embargo, lo más importante de todo es que el algoritmo tiene las mejores prestaciones con la función de coste de Kullback-Leibler, cuando antes esta función de coste no funcionaba en absoluto.

4.2. SIMULACIONES DEL ALGORITMO MODIFICADO

En estas simulaciones se comparan las prestaciones del algoritmo modificado frente al filtro FIR con umbrales adaptativos (Capítulo 2) y frente al algoritmo en escalera del capítulo 3. Las condiciones de los ensayos son las mismas que en los anteriores ensayos:

- ruido gaussiano blanco aditivo con una SNR de 25 y 30 dB;
- señal PAM de 8 niveles {±1, ±3, ±5, ±7}y símbolos uniformemente distribuidos;
- canal lineal de función de transferencia $H(z)=1+0,2z^{-1}$.

Se ensayan por separado las no linealidades con memoria y sin memoria. La no linealidad modelo es una saturación del tipo tangente hiperbólica con valor máximo 10.

Las funciones de coste comparadas son:

- Norma L_2 o error cuadrático con filtros de salida lineal;
- Norma L_2 o error cuadrático con filtros de salida tangente hiperbólica
- Norma entrópica

Se llevan a cabo los tres grupos de medidas hechas para los otros algoritmos:

- Tasa de error en recepción.
- Tiempo de convergencia.
- Capacidad de seguimiento de la variación del canal. de h[n]=δ[n]+0.2δ[n-1] a h[n]=δ[n]+0.05δ[n-1].

Recuérdese que la versión no discriminativa es aquella en la que el error que produce el dato a la salida se utiliza para entrenar todos los filtros independientemente de la cercanía o lejanía del dato a cada una de las fronteras de decisión. El entrenamiento discriminativo es aquél en el que se entrena sólo el filtro cuya frontera de decisión está más cerca del dato.

4.2.1. Versión no discriminativa

4.2.1.a No linealidad sin memoria

La Figura 4-1 muestra la tasa de fallos del algoritmo en escalera modificado y en su versión no discriminativa. Si se compara con la Figura 3-13,

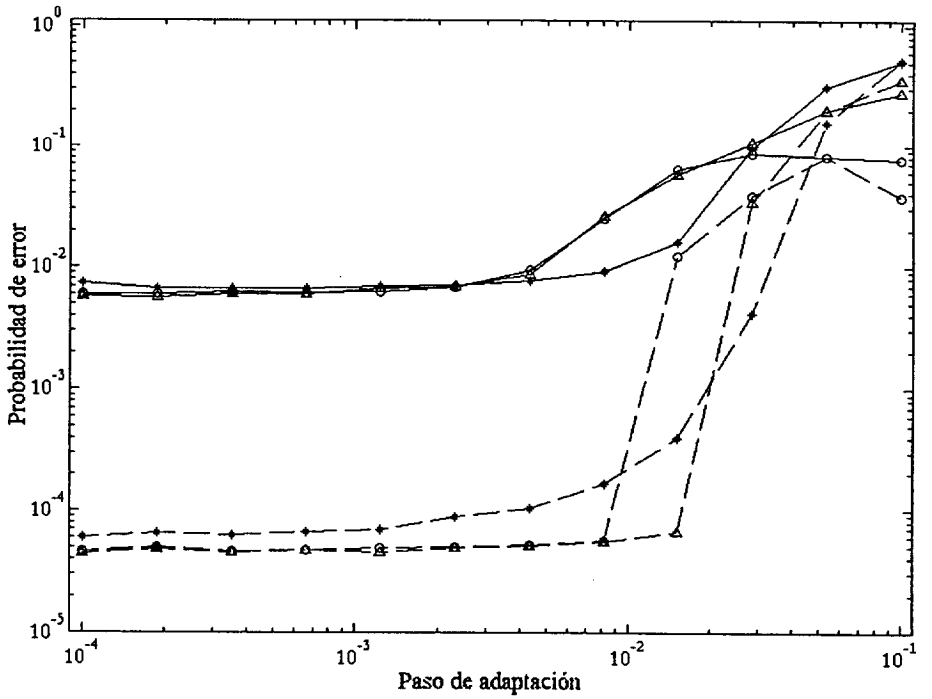


Figura 4-1. Tasa de error para el algoritmo en escalera modificado, versión no discriminativa. La no linealidad es sin menoria Las gráficas corresponden a coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (Δ) y coste entrópico (o). Las gráficas en trazo continuo corresponden a una SNR de 25 dB y las gráficas en trazo discontinuo a 30 dB. Se observa la mejora de las prestaciones con respecto al algoritmo original.

se observa la notable superioridad de este algoritmo con respecto al original. Las prestaciones son las mismas en el intervalo de paso de adaptación desde 10^{-4} a $6 \cdot 10^{-4}$, pero además la tasa de fallos se mantiene prácticamente constante hasta $0.8 \cdot 10^{-2}$ para el coste cuadrático con activación tangente hiperbólica y para un paso algo mayor en el coste entrópico: con la ventaja derivable de que un paso de adaptación mayor produzca una mayor velocidad de convergencia.

En la Figura 4-2 se observan los tiempos de convergencia para las tres funciones de coste y para distintos pasos de adaptación. La velocidad de convergencia del coste cuadrático con activación tangente hiperbólica alcanza a la del coste cuadrático con activación lineal prácticamente en todo el intervalo en el que la tasa de error es aceptable. Sin embargo, la velocidad de convergencia para las demás funciones de coste disminuye ligeramente.

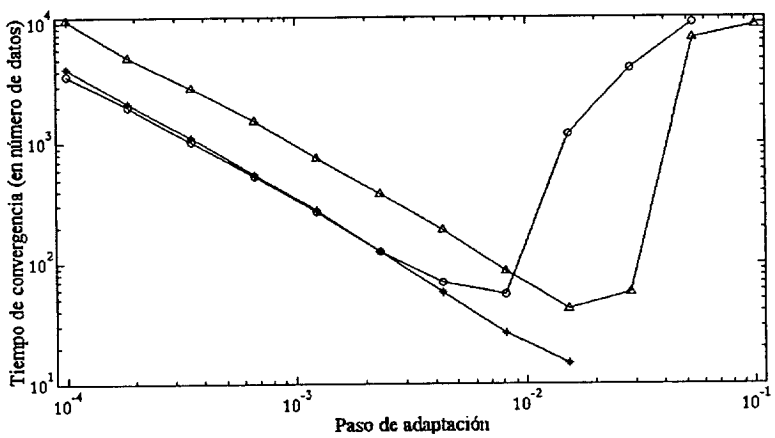


Figura 4-2. Tiempo de convergencia para el algoritmo en escalera modificado y en su versión no discriminativa. No linelidad sin memoria. Coste cuadrático con activación lineal (*), cuadrático con activación tangente hiperbólica (△) y entrópico (o).

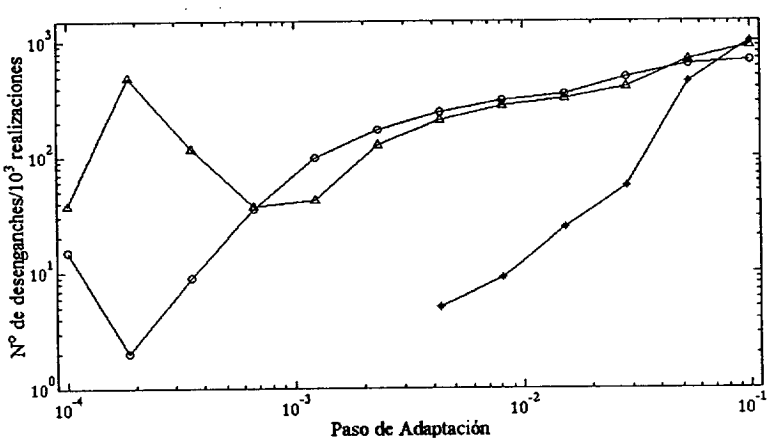


Figura 4-3. Número de desenganches en 1000 iteraciones para el algoritmo en escalera modificado y con la no linealidad a la entrada del canal. El coste cuadrático con activación lineal (*) es claramente superior a los costes cuadrático con activación tangente hiperbólica (△) y entrópico (o).

La robustez del coste cuadrático con activación lineal frente a cambios en el canal (Figura 4-3) es claramente superior a la de los otros dos costes. Sin embargo, el coste entrópico ofrece mayor velocidad de convergencia, porque se puede escoger un paso de adaptación cercano a 10^{-2} (con lo que converge en 100 muestras) manteniendo una buena tasa de error. El coste cuadrático hay que entrenarlo con pasos de adaptación menores y además la tasa de fallos se mantiene siempre ligeramente por encima de la del coste entrópico. Esto sugiere la posibilidad de combinar los costes de manera que cada uno de ellos actúe sólo cuando ofrece ventaja sobre los otros: aplicando una estrategia que haga que el algoritmo se guíe por el coste entrópico en el estado estacionario y que se utilice el coste cuadrático con activación lineal o una combinación de costes cuando el canal presente variaciones temporales bruscas.

4.2.1.b No linealidad con memoria

En la Figura 4-6 se observa una notable mejora de las prestaciones en cuanto a tasa de fallos respecto del algoritmo inicial. En cuanto al número de desenganches, el algoritmo con coste cuadrático y salida lineal es claramente

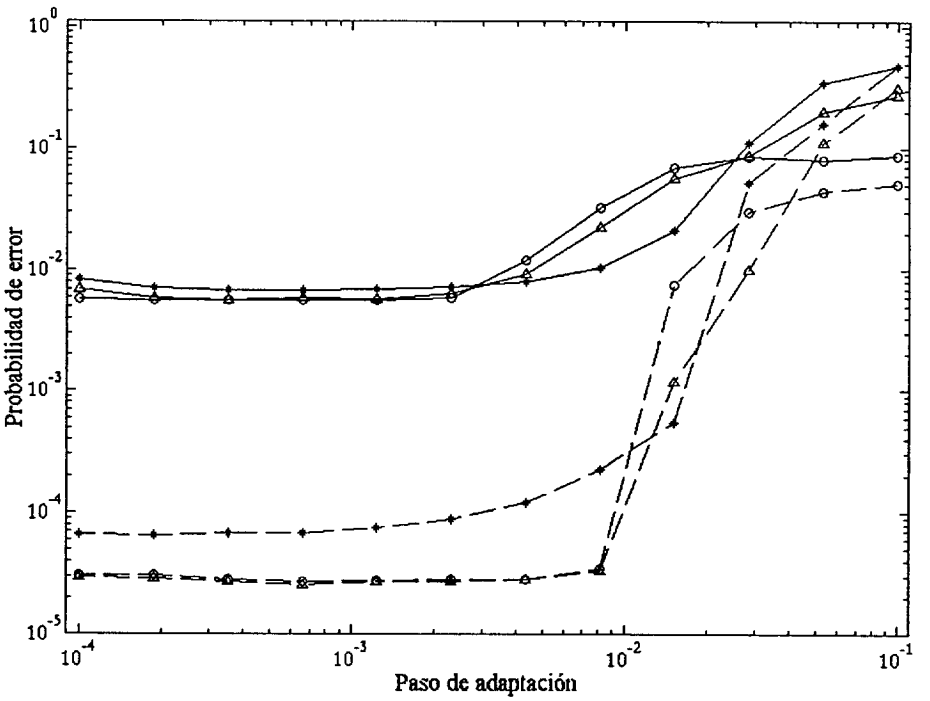


Figura 4-4. Tasa de error para el algoritmo en escalera modificado, versión no discriminativa. No linealidad con menoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o). Las gráficas en trazo continuo corresponden a una SNR de 25 dB y las gráficas en trazo discontinuo a 30 dB.

superior. Sin embargo, para pasos de adaptación menores a 10^{-3}, la tasa de

desenganches de las otras dos funciones de coste es menor al 1 por mil (cero en la gráfica). El tiempo de convergencia (Figura 4-5) del algoritmo con coste entrópico es comparable al tiempo de convergencia del algoritmo cuadrático con activación lineal.

Hay que destacar que, además de que el algoritmo converge con el coste entrópico, lo hace con la misma tasa de error que el cuadrático con sigmoide. La diferencia es más significativa cuando la no linealidad tiene memoria, porque los

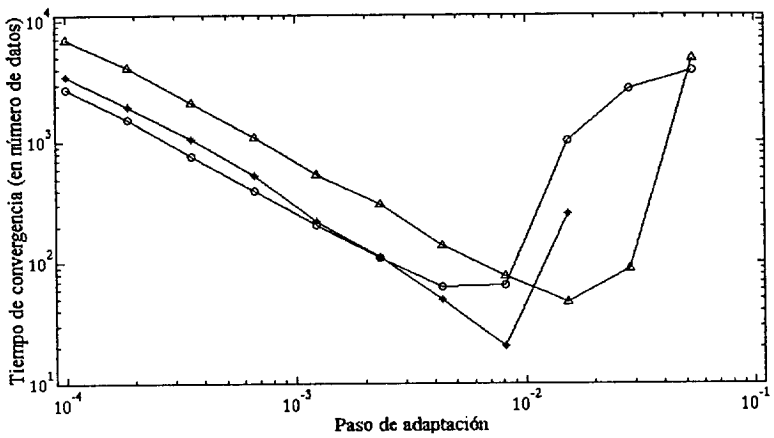


Figura 4-6. Tiempo de convergencia para el algoritmo en escalera modificado versión no discriminativa. No linealidad con memoria. Coste cuadrático con activación lineal (*), cuadrático con activación tangente hiperbólica (△) y entrópico (o).

datos están muy separados en algunas zonas y muy juntos en otras. Los costes con activación tangente hiperbólica no se ven afectados por los datos que están lejos al producir errores muy bajo. En cambio, el coste con activación lineal hace

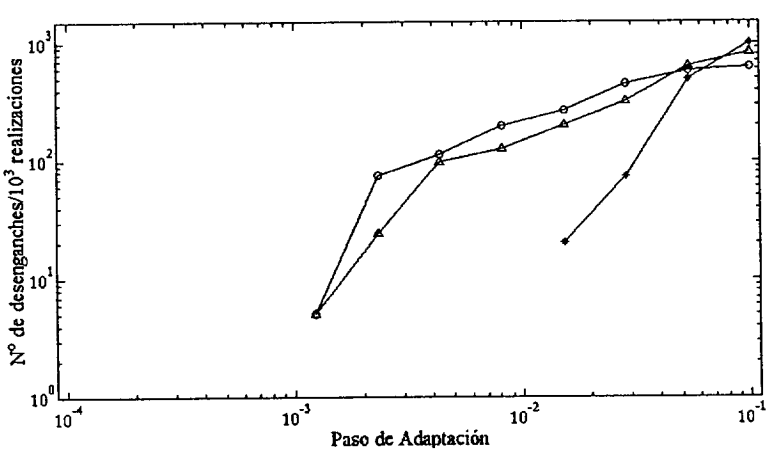


Figura 4-5. Número de desenganches en 1000 iteraciones para el algoritmo en escalera modificado no discriminativo. No linealidad con memoria. El coste cuadrático con activación lineal (*) es claramente superior a los costes cuadrático con activación tangente hiperbólica (△) y entrópico (o).

que la posición del hiperplano dependa no sólo de los datos cercanos sino de los datos que están lejos, que tienden a situar la frontera en una posición subóptima.

4.2.2. Versión discriminativa

Se observa que las prestaciones de esta versión no son peores que las de la versión no discriminativa, lo que justifica su uso al tener menor número de operaciones por dato (menor coste computacional).

4.2.2.a No linealidad sin memoria

Para la versión discriminativa (Figura 4-7) la tasa de error es similar a la de la versión no discriminativa. El tiempo de convergencia (Figura 4-8) aumenta ligeramente y el número de desenganches (Figura 4-9) disminuye con respecto a la versión no discriminativa.

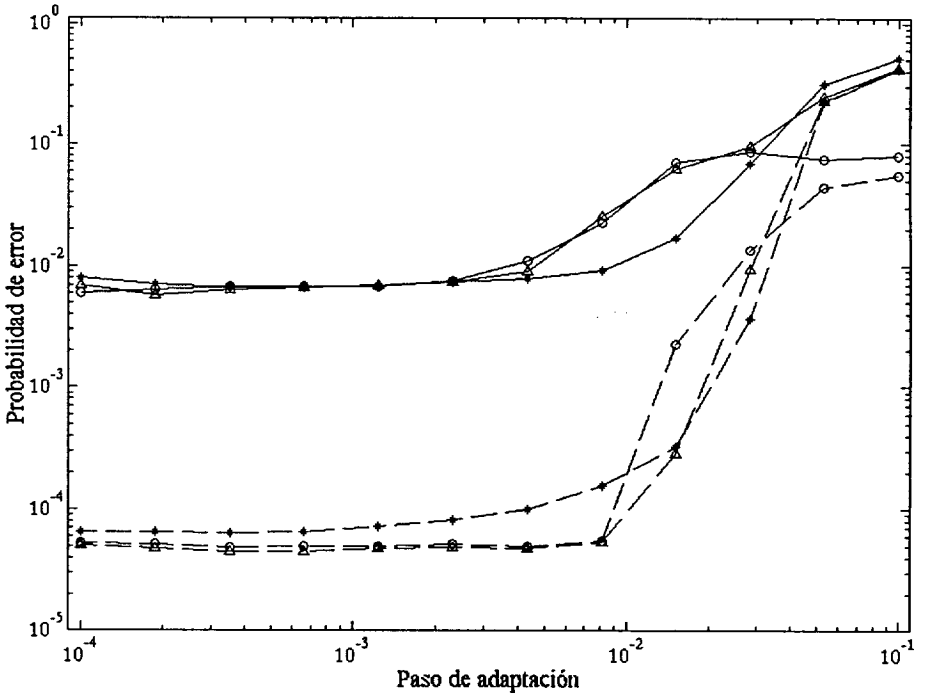


Figura 4-7. Tasa de error para el algoritmo en escalera modificado, versión discriminativa. No linealidad sin menoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o). Las gráficas en trazo continuo corresponden a una SNR de 25 dB y las gráficas en trazo discontinuo a 30 dB.

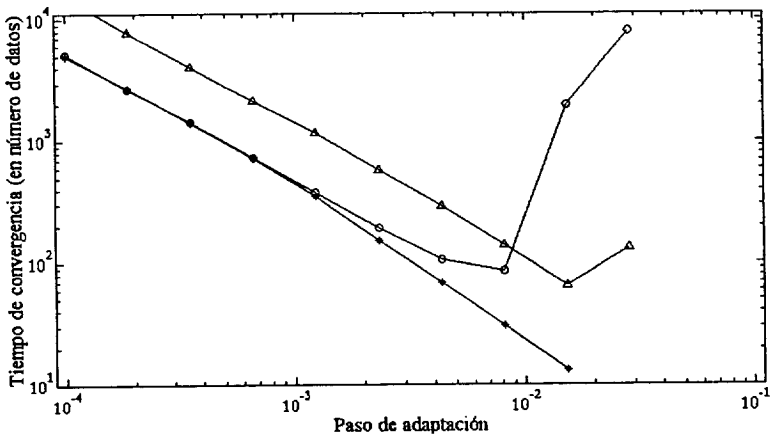


Figura 4-8. Tiempo de convergencia para el algoritmo en escalera modificado versión discriminativa. No linealidad sin memoria. Coste cuadrático con activación lineal (*), cuadrático con activación tangente hiperbólica (△) y entrópico (o).

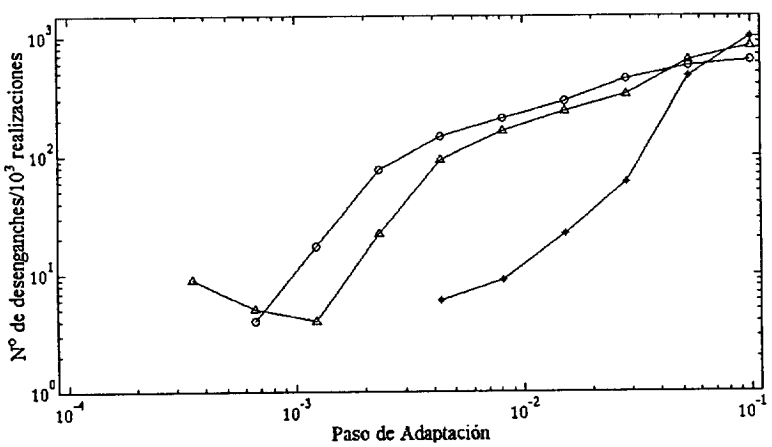


Figura 4-9. Número de desenganches en 1000 iteraciones para el algoritmo en escalera modificado discriminativo. No linealidad sin memoria. Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o). Nótese la mayor robustez del método discriminativo frente al no discriminativo.

4.2.2.b No linealidad con memoria

Las prestaciones de la versión discriminativa en tasa de error (Figura 4-10) son iguales a las de la versión no discriminativa cuando el canal presenta no linealidad con memoria. No se puede decir aquí tampoco que el algoritmo discriminativo degrade las prestaciones. La Figura 4-11 muestra el tiempo de convergencia del sistema. Aumenta un poco con respecto a la versión no discriminativa. El número de desenganches (Figura 4-12) disminuye.

La versión discriminativa produce tasas de error más bajas, ya que la posición de cada uno de los hiperplanos depende sólo de datos adyacentes a éstos. Sin embargo, esta mejora es muy pequeña, ya que hay otras causas que hacen que el error esté por encima del mínimo, como el ruido de desajuste.

En todo caso, el coste con activación tangente hiperbólica produce mejores resultados, al ser los datos que están más cerca del hiperplano los que modifican la posición de éste.

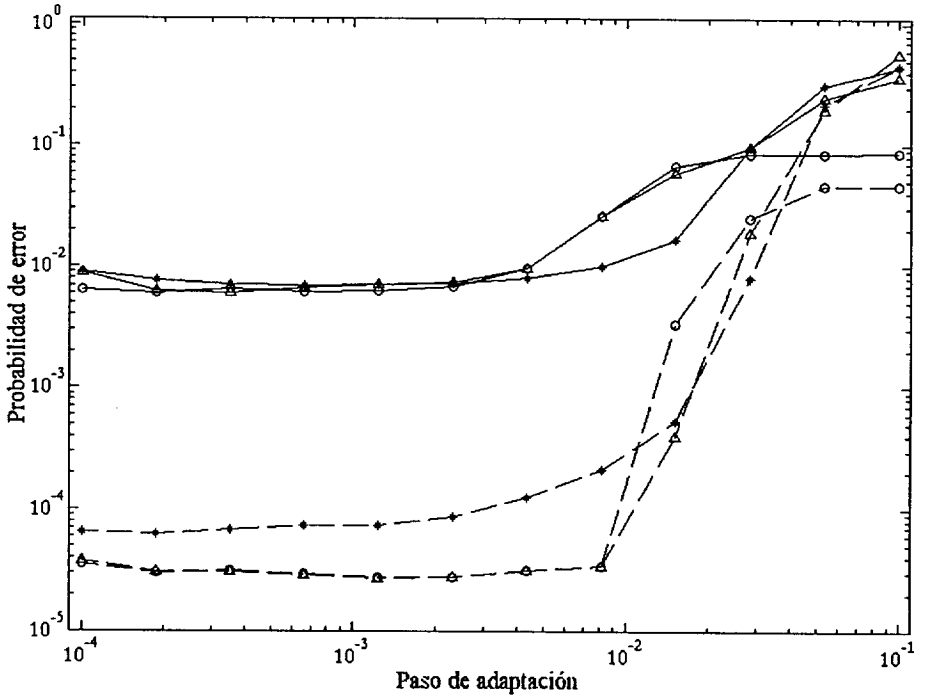


Figura 4-10. Tasa de error para el algoritmo en escalera modificado, versión discriminativa. No linealidad con menoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o). Las gráficas en trazo continuo corresponden a una SNR de 25 dB y las gráficas en trazo discontinuo a 30 dB.

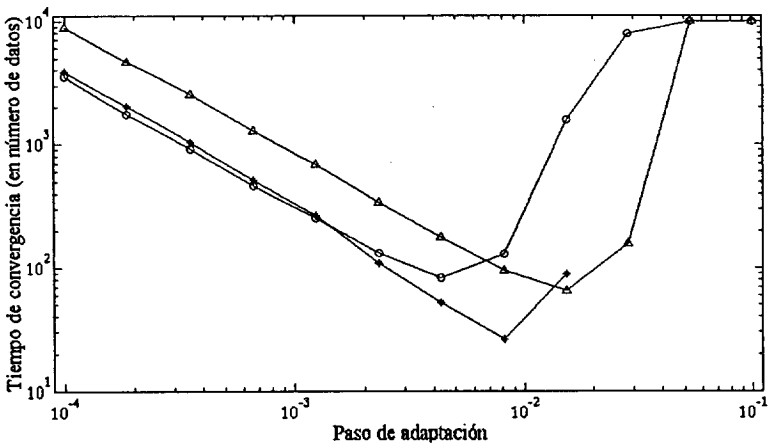


Figura 4-11. Tiempo de convergencia del algoritmo modificado, versión discriminativa. No linealidad con memoria Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o).

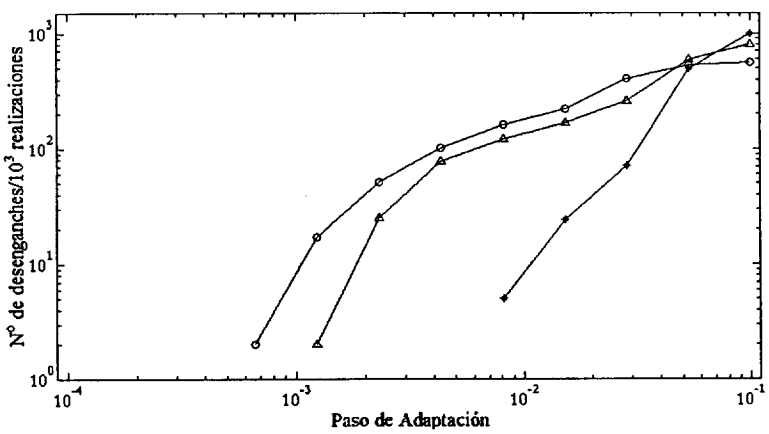


Figura 4-12. Número de desenganches en 1000 iteraciones para el algoritmo en escalera modificado discriminativo. No linealidad con memoria. Coste cuadrático con activación lineal (*), coste cuadrático con activación tangente hiperbólica (△) y coste entrópico (o).

4.3. DISCUSIÓN

El algoritmo modificado mejora el comportamiento en tasa de fallos del esquema en escalera porque extiende el intervalo en que converge con tasa de error pequeña: la tasa de fallos para el intervalo del paso de adaptación desde 10^{-4} a 10^{-3} disminuye hasta el mínimo valor que presentaba el algoritmo original.

El sistema presenta las mejores tasas de error para el coste entrópico, que antes no convergía.

El porcentaje de desenganches cuando el canal es no estacionario se reduce a valores menores al 1% para la variante discriminativa.

La velocidad de convergencia se reduce algo cuando la función de coste es cuadrática con activación tangente hiperbólica, aunque no varía para los demás costes.

Las tablas adjuntas muestran las prestaciones del algoritmo modificado respecto del original no discriminativo.

Tabla 4-1. Comparación de las tasas de fallo para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN). SNR = 30 dB. No linealidad sin memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	$6{,}6\cdot10^{-5}$	$5{,}0\cdot10^{-5}$	$5{,}6\cdot10^{-5}$	$8{,}0\cdot10^{-5}$	$1{,}5\cdot10^{-4}$	$1{,}9\cdot10^{-3}$
Esc. Mod, No disc.	$6{,}6\cdot10^{-5}$	$4{,}7\cdot10^{-5}$	$4{,}9\cdot10^{-5}$	$6{,}8\cdot10^{-5}$	$4{,}6\cdot10^{-5}$	$4{,}8\cdot10^{-5}$
Esc. Mod, Disc.	$6{,}6\cdot10^{-5}$	$4{,}7\cdot10^{-5}$	$5{,}24\cdot10^{-}$	$7{,}0\cdot10^{-5}$	$4{,}7\cdot10^{-5}$	$5{,}0\cdot10^{-5}$

Tabla 4-2. Comparación de las tasas de fallo para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN). No linealidad con memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	$6{,}7\cdot10^{-5}$	$3{,}0\cdot10^{-5}$	$3{,}4\cdot10^{-5}$	$9{,}1\cdot10^{-5}$	$1{,}0\cdot10^{-4}$	$8{,}8\cdot10^{-3}$
Esc. Mod, No disc.	$6{,}6\cdot10^{-5}$	$2{,}9\cdot10^{-5}$	$3{,}1\cdot10^{-5}$	$7{,}3\cdot10^{-5}$	$2{,}7\cdot10^{-5}$	$2{,}8\cdot10^{-5}$
Esc. Mod, Disc.	$6{,}4\cdot10^{-5}$	$3{,}2\cdot10^{-5}$	$3{,}2\cdot10^{-5}$	$7{,}4\cdot10^{-5}$	$2{,}8\cdot10^{-5}$	$2{,}8\cdot10^{-5}$

Tabla 4-3. Comparación de los tiempos de convergencia en número de muestras para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN). No linealidad sin memoria.

Paso de adaptación	$2\cdot10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	2160	1650	$>10^4$	620	480	-
Esc. Mod, No disc.	1980	4870	1890	356	975	344
Esc. Mod, Disc.	2540	7290	2570	456	144	479

Tabla 4-4. Comparación de las velocidades de convergencia para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN). No linealidad con memoria.

Paso de adaptación	$2 \cdot 10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	1850	1580	$>10^4$	540	470	-
Esc. Mod, No disc.	1840	3290	1450	300	692	259
Esc. Mod, Disc.	1915	4357	1657	335	852	315

Tabla 4-5. Comparación de los porcentajes de desenganche frente a canal no estacionario para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN). No linealidad sin memoria.

Paso de adaptación	$2 \cdot 10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	0	76	100	0	79	10
Esc. Mod, No disc.	0	43	0,2	0	4	6,8
Esc. Mod, Disc.	0	0	0	0	0,4	1

Tabla 4-6. Comparación de los porcentajes de desenganche frente a canal no estacionario para las funciones de coste cuadrática con activación lineal (CL), cuadrática con activación tangente hiperbólica (CT) y entrópica (EN). No linealidad sin memoria.

Paso de adaptación	$2 \cdot 10^{-4}$			10^{-3}		
Función de Coste	CL	CT	EN	CL	CT	EN
Escalera No disc	0	2,6	100	0	24	100
Esc. Mod, No disc.	0	0	0	0	0	0
Esc. Mod, Disc.	0	0	0	0	0	0,8

Las tablas 4-1 y 4-2 muestran la comparación de tasas de error. El coste entrópico presenta las mejores tasas, sin degradación significativa en el algoritmo discriminativo. En las tablas 4-3 y 4-4 se observa en particular cómo el coste entrópico tiene mayor velocidad de convergencia con respecto al algoritmo en escalera no modificado. En cuanto a porcentaje de desenganche, en las tablas 4-4 y 4-5 se puede observar que el coste entrópico presenta cero desenganches sobre mil simulaciones para un paso de adaptación de $2 \cdot 10^{-4}$ cuando en el algoritmo original fallaba siempre. Si el paso de adaptación aumenta a 10^{-3}, el numero de desenganches está en torno al 1%. Esta tasa aumenta con el paso de adaptación, mateniéndose más baja con el coste cuadrático con activación lineal. Si se quiere aumentar la velocidad de convergencia hasta el máximo permitido por la tasa de errores, la estabilidad frente a canales no estacionarios es baja con el coste entrópico, en tanto que en el coste cuadrático con activación lineal es elevada. Por el contrario, la tasa de fallos del coste cuadrático con activación lineal es elevada con pasos de adaptación grandes. Por ello se sugiere la posibilidad de

combinar ambos costes.

ANEXO 4.1 OTROS ESQUEMAS

A4.1.1 Esquema alternativo para el algoritmo en escalera

El algoritmo en escalera consiste en una cadena de filtros, todos ellos del mismo orden. En caso de distorsión no lineal, los datos de los extremos estarán mucho más juntos entre sí que los datos del interior, de modo que puede que sea necesario aumentar el orden de los filtros exteriores. Para ello se pueden aprovechar de las capas anteriores: en lugar de introducir los datos de entrada en la capa siguiente, se pueden usar los datos pasados por el filtro anterior, más un dato adicional. El esquema puede verse en la Figura 4-13.

La primera capa consiste en un filtro FIR de orden 2 que separa los datos correspondientes a símbolos positivos de los correspondientes a símbolos negativos. El signo de su salida se utiliza para seleccionar qué elementos de la siguiente capa serán seleccionados y para determinar el valor del bit más significativo.

La siguiente capa toma la salida de la anterior más una muestra adicional de la señal. Es, por tanto, un filtro FIR de orden 3. El signo de la primera salida condiciona el valor del peso independiente de este filtro: si el dato produce una

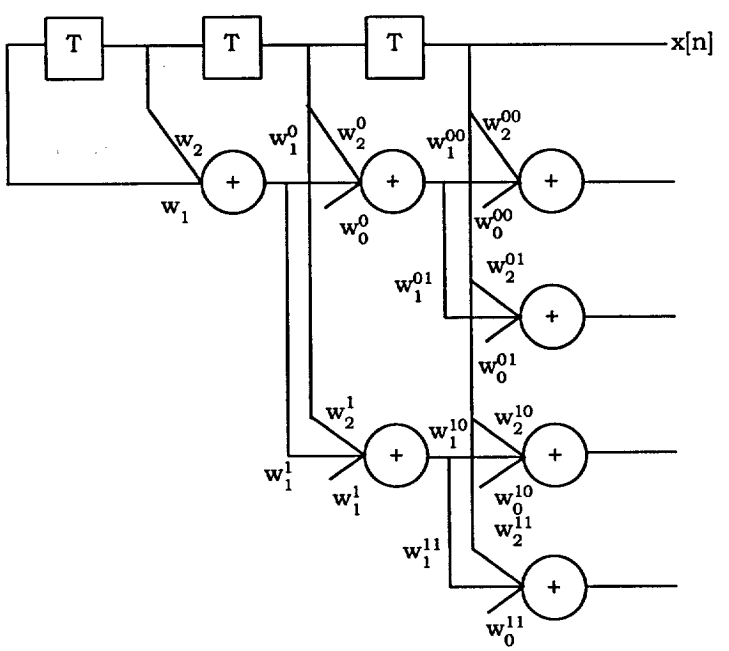


Figura 4-13. Algoritmo en escalera modificado. La salida del filtro de la capa 0, de orden 2, se aplica a las entradas de los filtros de la capa 1, formando un filtro de orden 3, etc.

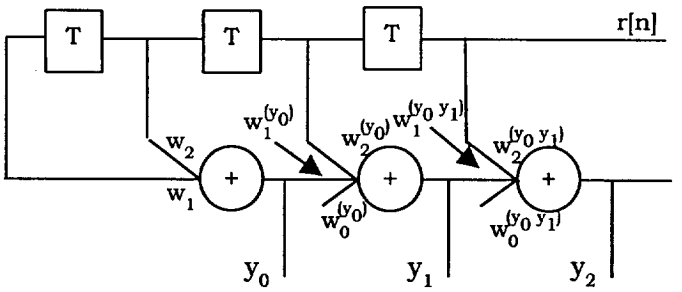


Figura 4-14. Notación simplificada para el igualador. La primera salida corresponde a la primera capa del algoritmo en escalera. Esta salida, además de ser aplicada al segundo filtro, condiciona el valor del peso independiente de éste, y así sucesivamente hasta llegar al último filtro. Cada una de las salidas determina uno de los bits del símbolo, igual que en el algoritmo en escalera original.

salida negativa en el filtro anterior, se elige el signo del peso independiente de manera que el plano de separación del presente filtro esté situado en la zona negativa del espacio, y viceversa.

Se procede así de manera iterativa hasta llegar a la última capa. Esto es equivalente al filtro de la Figura 4-14, donde los pesos de cada sumador dependen de la salida del sumador anterior.

Si es necesario hacer que el filtro de entrada tenga un orden superior, todos los demás filtros aumentarán también su orden, sin necesidad de modificarlos.

Para una constelación de P símbolos no es necesario disponer P/2 filtros de orden M, sino que se dispone de un filtro de orden M-L-1 (siendo $P=2^L$) seguido de P/2-1 capas de filtros de orden 2 conectados en la forma vista.

A4.1.2 Extensión del algoritmo a constelaciones complejas

Las modulaciones complejas, en particular la QAM, son necesarias cuando es necesario transmitir una gran cantidad de información por un canal de comunicaciones de ancho de banda limitado. Cuando se hace esto, el diseñador cambia ancho de banda por potencia transmitida: cuantos más símbolos tenga la constelación, peores serán las prestaciones en cuanto a relación señal a ruido y, por lo tanto, la potencia transmitida deberá ser mayor. Es entonces cuando no se puede abusar en exceso del "back-off" del emisor ya que aparecerán fenómenos no lineales importantes.

El algoritmo que se presenta aquí tiene una inmediata extensión a QAM. Para $M=L^2$ símbolos en cuadratura son necesarias $(L-1)^2$ superficies de discriminación. Sin embargo, la constelación tendrá simetría respecto de los dos ejes real e imaginario, de manera que sólo es necesario mantener las rectas de discriminación que se hallen en un cuadrante, es decir, $L^2/4$ rectas (Figura 4-15).

El primer filtro detecta los signos de las componentes en fase y en cuadratura. Para los siguientes filtrados deben, o bien eliminarse los signos de ambas componentes para que pasen a ocupar el primer cuadrante y restaurarse después, o bien girar los filtros 0, 90, 180 o 270° dependiendo del cuadrante en el que se halle el dato.

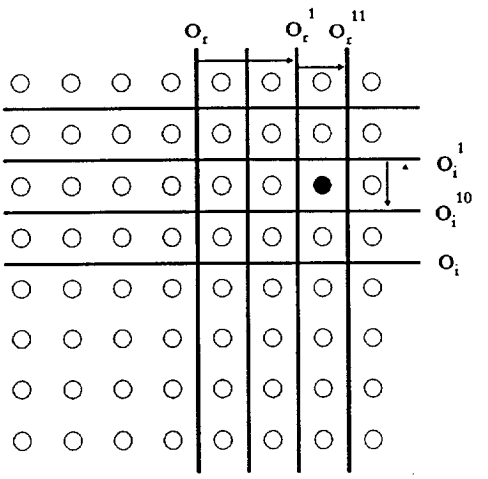


Figura 4-15. Secuencia de decisiones en una constelación 64-QAM.

A4.1.3 Igualador en escalera visto como cadena de expertos

El algoritmo en escalera puede verse como una estructura modular, donde cada filtro tiene por supervisor el filtro anterior. El esquema equivalente está en la Figura 4-16, para el caso particular de señales 8-PAM.

Visto de esta manera, el igualador es una cadena de decisiones "bit a bit". Cada experto tiene una salida binaria y su complementario. El primer experto decide el bit de signo b_0. La segunda capa, con dos expertos, decide el bit b_1, el resultado del cual se multiplica por el resultado del bit anterior, y así sucesivamente, las capas siguientes decidirán cada una un bit cada vez menos significativo hasta llegar al último. Para el caso de señales P-PAM, son necesarios L bits, donde $P=2^L$ y, por lo tanto, L capas. El número de expertos necesarios es $P-1=2^L-1$. Sólo habrá una salida no nula, que corresponderá a la decisión sobre el símbolo emitido.

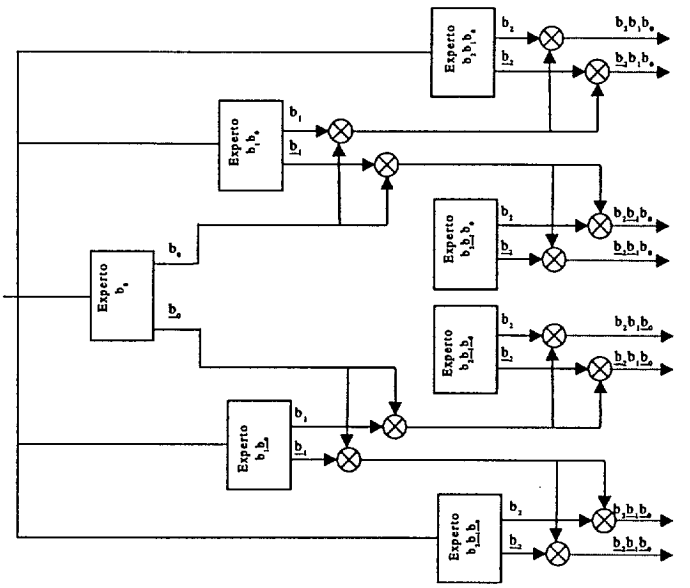


Figura 4-16. Igualador en escalera visto como un conjunto de expertos, para recepción 8-PAM.

Cada uno de los expertos de cada una de las capas se especializa en una zona diferenciada, pero, a diferencia del caso de las redes modulares, este sistema se basa en capas con sucesivos niveles de especialización, como ya ha sido expuesto más arriba. En el apartado 1.1 se verá una manera de extraer la información de una capa para acelerar el entrenamiento de la capa siguiente. Ello redunda en un mayor aprovechamiento de los datos a fin de que el entrenamiento de un experto no esté limitado exclusivamente a los datos de su entorno.

ANEXO 4.2: ANÁLISIS DEL ALGORITMO MODIFICADO

A4.2.1 Filtro transversal óptimo de Wiener

Sea un conjunto de vectores de datos de dimensión M+1. El dato n-ésimo tiene la forma $\mathbf{z}[n]=[z_0[n],\cdots,z_{M-1}[n],1]$. Para cada dato se calcula su producto escalar con el vector de coeficientes $\mathbf{w}=[w_1,\cdots,w_M]$ (filtro FIR) para formar la salida o[n]:

$$o[n]=\mathbf{z}^T[n]\cdot\mathbf{w} \tag{4-6}$$

Sea d_n la salida deseada correspondiente al dato z[n]. Los coeficientes se mantienen constantes y los datos, así como la salida deseada, son procesos ergódicos y estacionarios. El error cuadrático correspondiente al dato z[n] es la cantidad

$$\varepsilon^2[n]=(d_n-o[n])^2=(d_n-\mathbf{z}^T[n]\cdot\mathbf{w})^2 \tag{4-7}$$

Se llama error cuadrático medio a la cantidad

$$\begin{aligned}E\{\varepsilon^2[n]\}&=E\{(d_n-\mathbf{z}^H[n]\cdot\mathbf{w})^2\}\\&=E\{d_n^2\}-2E\{d_n\cdot\mathbf{z}^H[n]\}\mathbf{w}+\mathbf{w}^HE\{\mathbf{z}[n]\cdot\mathbf{z}^H[n]\}\\&=E\{d_n^2\}-2\mathbf{pw}+\mathbf{w}^H\mathbf{Rw}\end{aligned} \tag{4-8}$$

donde el vector de correlación cruzada de los datos con la salida deseada se define como

$$\mathbf{p}=E\{d_k\mathbf{z}^*[n]\}=E\begin{bmatrix}d_kz_1^*[n]\\\vdots\\d_kz_{M-1}^*[n]\\d_k\end{bmatrix} \tag{4-9}$$

y la matriz de autocorrelación de los datos se define como

$$\mathbf{R}=E\{\mathbf{z}[n]\cdot\mathbf{z}^H[n]\}=E\begin{bmatrix}z_1[n]\cdot z_1^*[n] & z_1[n]\cdot z_2^*[n] & \cdots\\z_2[n]\cdot z_1^*[n] & & \\\vdots & & \end{bmatrix} \tag{4-10}$$

Para calcular el vector $\mathbf{w}^*$ que proporciona el mínimo error cuadrático medio, se

calcula el gradiente de éste respecto de **w** en la ecuación (4-8):

$$\nabla_w \equiv -2\mathbf{p} + 2\mathbf{R}\mathbf{w} \tag{4-11}$$

Igualando el gradiente a cero se obtiene el vector óptimo [Widrow, 1995]

$$\mathbf{w}^* = \mathbf{R}^{-1}\,\mathbf{p} \tag{4-12}$$

La ecuación (4-20) es la forma matricial de la ecuación de Wiener-Hopf.

El algoritmo de máxima pendiente varía el valor del vector de coeficientes (inicializado de cierta manera) mediante pequeños cambios en la dirección del gradiente y en sentido contrario:

$$\mathbf{w}[n] = \mathbf{w}[n-1] - \mu\nabla_w \zeta \tag{4-13}$$

Una versión del algoritmo de máxima pendiente es el Least Mean Squares (LMS). La estimación del gradiente respecto de **w** que utiliza este algoritmo es el valor instantáneo del error en cada instante

$$\varepsilon^2 = \left(d_n - \mathbf{z}^T[n]\cdot\mathbf{w}\right)^2 \tag{4-14}$$

y la actualización del vector de pesos es

$$\mathbf{w}[n] = \mathbf{w}[n-1] + 2\mu\,\varepsilon[n]\mathbf{z}[n] \tag{4-15}$$

Para un valor fijo de **w** este algoritmo es insesgado. En efecto:

$$E\{\hat{\zeta}\} = -2E\{\,\varepsilon[n]\mathbf{z}[n]\,\} = -2E\{\,d_k\,\mathbf{z}[n] - \mathbf{z}[n]\,\mathbf{z}^T[n]\mathbf{w}\} = -2\mathbf{p} + 2 \tag{4-16}$$

que es la solución de Wiener.

El algoritmo LMS converge en media y con varianza pequeña a la solución de Wiener siempre que el paso de adaptación μ sea suficientemente pequeño. Una cota superior de μ es

$$\mu < \frac{1}{\lambda_{max}} \tag{4-17}$$

siendo λ_{max} el mayor autovalor de la matriz de autocorrelación. Esta cota asegura la convergencia en media, pero no asegura varianza pequeña. Otra cota superior que asegura la convergencia en varianza es [Widrow, 1992]

$$\mu < \frac{1}{\mathrm{tr}\,\mathbf{R}} \tag{4-18}$$

A4.2.2 Notación

A efectos del análisis, se toma la siguiente notación. Llamamos $\mathbf{r}^{b_0\cdots b_j}[n]$ al conjunto de vectores de entrada al filtro $o^{b_0\cdots b_j}$. En el caso de que este filtro sea el de la primera capa, $\mathbf{r}[n]$ comprende a las P clases de datos; en el caso en que el filtro sea de la segunda capa, las clases de datos comprendidas en el

conjunto de vectores $\mathbf{r}^{b_0}[n]$serán sólo la mitad, y así sucesivamente hasta llegar a la última capa de filtros, donde las entradas a cada uno de los filtros corresponden tan solo a dos clases contiguas. El filtro que debe utilizarse está determinado por el valor $x[n]$ del primer componente del vector de datos $\mathbf{r}^{b_0 \cdots b_j}[n]$. Así, para el único filtro de la primera capa, $x[n]$ toma cualquiera de los valores del alfabeto.

En un canal canal lineal, y en ausencia de ruido, la entrada $\mathbf{r}^{b_0 \cdots b_j}[n]$ al receptor tiene la siguiente expresión:

$$\mathbf{r}^{b_0 \cdots b_j}[n] = \begin{bmatrix} x^{b_0 \cdots b_j}[n] & \cdots & x[n-N+1] \\ \vdots & & \vdots \\ x[n-M+1] & \cdots & x[n-N-M+2] \end{bmatrix} \mathbf{h} = \mathbf{X}^{b_0 \cdots b_j}[n]\mathbf{h} \tag{4-19}$$

Es decir, en la matriz de símbolos $\mathbf{X}^{b_0 \cdots b_j}[n]$de la ecuación (4-19) el elemento $x^{b_0 \cdots b_j}[n]$ pertenece necesariamente a la clase de símbolos cuyos j+1 bits más significativos tienen los valores $b_0 \ldots b_j$. Los demás elementos x[n-k], k>0 son símbolos pertenecientes a cualquiera de las clases.

En el caso en que la no linealidad esté en la entrada del canal (no linealidad sin memoria), la matriz de datos tendrá los elementos F(x[n-k]) y es válido lo anterior.

Para el caso de no linealidad con memoria no es posible definir una matriz de símbolos de entrada al receptor, puesto que la entrada es una función no lineal con la forma

$$G\left(\mathbf{x}^t[n]\mathbf{h}\right) \tag{4-20}$$

Sin embargo, el análisis basado en expansión de Volterra permite expresar la función como una serie de funciones lineales y, en esta expansión se utilizará la matriz de símbolos.

Por otro lado, $\mathbf{w}^{b_0 \cdots b_j}$ es el vector de pesos del filtro cuya entrada es el vector

$$\left(\mathbf{r}^{b_0 \cdots b_j}[n]\,1\right)^T$$

El término constante se corresponde con el término independiente del vector de pesos.

Todo el análisis se lleva a cabo sin tener en cuenta el ruido y suponiendo que los filtros del igualador cumplen la condición (4-12) (filtros óptimos) y que no hay errores en la decisión que puedan afectar a la distribución estadística de

los datos de entrada a cada uno de los filtros siguientes, esto es, se supone que los datos de entrada a cada uno de los filtros son los que se desean.

A4.2.3 Entrenamiento modificado en el caso de canal lineal

Cuando el canal es lineal y los datos están idénticamente distribuidos, los filtros serán exactamente iguales, excepto por su posición en el espacio de datos. Es decir, las ecuaciones de los filtros representan una serie de espacios afines paralelos al plano que representa la ecuación del filtro central. Esto es lo que pretendemos mostrar en este apartado.

En efecto, el vector óptimo de pesos $\mathbf{w}^{*b_0 \cdots b_i}$ cumple la condición

$$\mathbf{R}^{b_0 \cdots b_i} \mathbf{w}^{*b_0 \cdots b_i} = \mathbf{p}^{b_0 \cdots b_i} \tag{4-21}$$

donde $\mathbf{R}^{b_0 \cdots b_i}$ es la matriz de autocorrelación de los datos correspondientes al conjunto de entrada al filtro $\mathbf{w}^{*b_0 \cdots b_i}$ y $\mathbf{p}^{b_0 \cdots b_i}$ es el vector de correlación cruzada entre los datos y la señal deseada. Las expresiones de generales de $\mathbf{R}$ y $\mathbf{p}$ para cualquiera de los filtros para canal lineal son

$$\mathbf{R} = E\left\{ \begin{bmatrix} \mathbf{r}[n] \\ 1 \end{bmatrix} \cdot \begin{bmatrix} \mathbf{r}[n] \\ 1 \end{bmatrix}^T \right\}^T =$$

$$= \begin{bmatrix} \rho^2[0,0] & \rho^2[0,1] & \cdots & \rho^2[0,M-1] & \rho^1[0] \\ \rho^2[1,0] & \ddots & & & 0 \\ \vdots & & \rho^2[p,q] & & \vdots \\ \rho^2[M-1,0] & & & \ddots & 0 \\ \rho^1[0] & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{4-22}$$

$$\mathbf{p} = E\{\mathbf{d}\,\mathbf{r}[n]\} = \begin{bmatrix} E\{d[n] \cdot r[n]\} \\ \vdots \\ E\{d[n] \cdot r[n-M+1]\} \\ E\{d[n]\} \end{bmatrix} = \begin{bmatrix} p \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4-23}$$

Los elementos $\rho^2[p,q]$ de la matriz **R** tienen la expresión siguiente:

$$\rho^2[p,q] = E\{r[n-p]\cdot r^*[n-q]\} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} h_i h_j^* E\{x[n-p-i]x[n-q-j]\} \quad (4\text{-}24)$$

Suponiendo $0 \le p,q < M$ y sabiendo que los símbolos x[n-k] son independientes e idénticamente distribuidos, la ecuación anterior toma la siguiente forma para p+q>0

$$\rho^2[p,q] = \sum_{j=p-q}^{N-1} h_{q-p+j} h_j^* E\{x^2[n-q-j]\} = \quad 0 > p-q \ge N-1$$
$$= \sigma_x^2 \sum_{j=p-q}^{N-1} h_{q-p+j} h_j^* = \sigma_x^2 \rho_h^2[p-q] \quad (4\text{-}25)$$

$$\rho^2[p,q] = 0, \qquad p-q > N-1$$

siendo σ_x^2 la varianza de los símbolos. Estos momentos son iguales para todos los filtros. $\rho_h^2[p,q]$ y σ_x^2 no dependen del filtro al cual corresponde la matriz de autocorrelación anterior.

Para p=q=0 la situación es diferente, ya que interviene x[n] en el sumatorio (4-24). Como x[n] corresponde a un conjunto de valores diferente para cada uno de los filtros, su estadística varía. Por esto, para el elemento en p=q=0 utilizamos la notación $\rho^2_{\mathbf{r}\,b_0\ldots b_j}[0,0]$.

La ecuación (4-24) particularizada a este caso es:

$$\rho^2_{\mathbf{r}\,b_0\ldots b_j}[0,0] = \sum_{i=p-q}^{N-1} \|h_i\|^2 E\left\{\left(x^{b_0\ldots b_j}[n-i]\right)^2\right\} = \quad (4\text{-}26)$$
$$= \|\mathbf{h}\|^2 \sigma_x^2 + h_0^2\left(m^2_{X\,b_0\ldots b_j} + \sigma^2_{X\,b_0\ldots b_j} - \sigma_x^2\right)$$

Finalmente, es fácil ver que

$$\rho^1[0] = \rho^1_{\mathbf{r}\,b_0\ldots b_j} = h_o m_{X\,b_0\ldots b_j} \quad (4\text{-}27)$$

Con esto, la ecuación (4-22) queda en la forma

$$\mathbf{R}^{b_0 \ldots b_j} = \begin{bmatrix} \|\mathbf{h}\|^2 \sigma_x^2 + h_0^2\left(m_X^2{}^{b_0 \ldots b_j} + \sigma_X^2{}^{b_0 \ldots b_j} - \sigma_x^2\right) & \rho_h^2[0,1]\sigma_x^2 & \cdots & \rho_h^2[0,M-1]\sigma_x^2 & h_0 m_X{}^{b_0 \ldots b_j} \\ \rho_h^2[1,0]\sigma_x^2 & \ddots & & & \\ \vdots & & \ddots & & \vdots \\ \rho_h^2[M-1,0]\sigma_x^2 & & & \ddots & 0 \\ h_0 m_X{}^{b_0 \ldots b_j} & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{4-28}$$

Sustituyendo la anterior matriz en la ecuación matricial $\mathbf{R}^{b_0 \ldots b_j}\mathbf{w}^{*b_0 \ldots b_j} = \mathbf{p}^{b_0 \ldots b_j}$ y despejando la última fila se obtiene[17]

$$w_M^{b_0 \ldots b_j} = h_0 m_X{}^{b_0 \ldots b_j} \; w_0^{b_0 \ldots b_j} \tag{4-29}$$

Si se despeja $w_M^{b_0 \ldots b_j}$ se obtiene el sistema de M ecuaciones

$$\begin{bmatrix} \|\mathbf{h}\|^2 \sigma_x^2 + h_0^2\left(\sigma_X^2{}^{b_0 \ldots b_j} - \sigma_x^2\right) & \rho_h^2[1,0]\sigma_x^2 & \cdots & \rho_h^2[M-1,0]\sigma_x^2 \\ \rho_h^2[0,1]\sigma_x^2 & \ddots & \vdots & \\ \vdots & & & \\ \rho_h^2[0,M-1]\sigma_x^2 & \cdots & & \end{bmatrix} \cdot \begin{bmatrix} w_0^{b_0 \ldots b_j} \\ \vdots \\ \vdots \\ w_{M-1}^{b_0 \ldots b_j} \end{bmatrix} = \begin{bmatrix} p \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4-30}$$

El sistema de ecuaciones descrito por las filas 2 a M es un sistema de M-1 ecuaciones y M-1 incógnitas que describe una línea que pasa por el origen de coordenadas. Esta línea tiene un vector director $\left(w_1^{b_0 \ldots b_j}, \cdots, w_{M-1}^{b_0 \ldots b_j}\right)^T$ cuya dirección es independiente de los símbolos $x^{b_0 \ldots b_j}$, (su varianza no aparece entre las ecuaciones): esta recta es la misma cualquiera que sea el clasificador al que pertenezca la matriz de autocorrelación (4-28). La primera ecuación es un hiperplano en el espacio R^M, que no pasa por el origen y cuya posición depende de $x^{b_0 \ldots b_j}$. Su intersección con la recta anteriomente descrita produce la solución al sistema de ecuaciones (4-30). Por lo tanto, todos los filtros tendrán soluciones

[17] El término independiente es proporcional a la media de los símbolos de entrada al filtro $b_0 \ldots b_j$, lo cual ya ha sido comprobado anteriormente y además es la solución que razonablemente se espera: el plano de discriminación descrito por la función de transferencia del filtro corta al eje principal del espacio de datos $r[n]$ en $r = h_0 E\left\{x^{b_0 \ldots b_j}[n]\right\}$.

al vector $\mathbf{w}^{b_0 \dots b_j}$ con la misma dirección en el espacio.

Dos conclusiones se extraen inmediatamente del anterior resultado:

1. Las superficies de discriminación serán paralelas.

En efecto, llamando $\mathbf{w}^*$ a la solución para el filtro del clasificador central, los vectores de coeficientes de todos los filtros serán proporcionales al vector

$$\hat{\mathbf{w}}^{*b_0 \dots b_j} = \frac{\mathbf{w}^{*b_0 \dots b_j}}{w^{*b_0 \dots b_j}_0} = \left[1 \quad \frac{w^{*b_0 \dots b_j}_1}{w^{*b_0 \dots b_j}_0} \quad \frac{w^{*b_0 \dots b_j}_{M-1}}{w^{*b_0 \dots b_j}_0} \quad \cdots \quad h_0 m_{x^{b_0 \dots b_j}} \right]$$

(Recuérdese que $w^{b_0 \dots b_j}_M = h_o m_{x^{b_0 \dots b_j}} \; w^{b_0 \dots b_j}_0$ en la solución)

Como los elementos 1 a M-1 del anterior vector de M elementos serán iguales para todos los filtros, se puede escribir:

$$\hat{\mathbf{w}}^{*b_0 \dots b_j} = \frac{\mathbf{w}^{*b_0 \dots b_j}}{w^{*b_0 \dots b_j}_0} = \left[1 \quad \frac{w^*_1}{w^{*b_0 \dots b_j}_0} \quad \frac{w^*_{M-1}}{w^{*b_0 \dots b_j}_0} \quad \cdots \quad h_0 m_{x^{b_0 \dots b_j}} \right] \tag{4-31}$$

La posición relativa de los planos viene determinada únicamente por el término independiente $\hat{w}^{*b_0 \dots b_j}_M = h_0 m_{x^{b_0 \dots b_j}}$.

2. La posición de las superficies de separación no depende de si el entrenamiento es o no discriminativo

Esto se deduce observando el mismo aspecto de la matriz: cualquiera que sea el conjunto de datos utilizado para formar la matriz de autocorrelación (todos los datos que se introducen en el filtro o bien sólo los de las dos clases más cercanas a la superficie), los valores de las filas 2 a M son iguales. Las únicas filas que varían son la primera y la última, de cuyo valor depende el módulo del vector de pesos del filtro.

Como para la clasificación sólo se requiere el signo de la salida, una solución equivalente para todos los filtros es la solución del filtro de la primera capa, excepto para el término independiente, que se debe calcular por separado: una modificación del algoritmo en escalera para igualación en canal lineal pasa por entrenar el filtro principal (cuyo término independiente debe forzarse a cero, ya que la media de los datos es nula) y pasar el resultado del entrenamiento a todos los demás filtros.

El término independiente se calcula para cada filtro sólo con los datos

correspondientes a ese filtro. Todo lo que hay que calcular es su media o, equivalentemente, entrenar ese peso mediante LMS[18]. Todo ello disminuye significativamente el coste computacional.

De esta manera se consigue, además, que el sistema sea prácticamente tan estable respecto del paso de adaptación como lo es un solo filtro entrenado mediante LMS. Eso soluciona el problema de la inestabilidad del coste de Kullback-Leibler en el caso particular de canal lineal.

A4.2.4 Entrenamiento en canal con no linealidad sin memoria

Cuando la no linealidad está en la entrada del canal, entonces el vector de datos en ausencia de ruido tiene la forma

$$\mathbf{r}^{b_0 \cdots b_i}[n] = \begin{bmatrix} F\left(x^{b_0 \cdots b_i}[n]\right) & \cdots & F(x[n-N+1]) \\ \vdots & & \vdots \\ F(x[n-M+1]) & \cdots & F(x[n-N-M+2]) \end{bmatrix} \mathbf{h} \qquad (4\text{-}32)$$

Basta con substituir los símbolos emitidos x[n-k] por F(x[n-k]) en las ecuaciones anteriores. Si bien los elementos $r^{b_0 \cdots b_i}[n]$ son diferentes a los del caso lineal, es fácil ver que el resultado tan solo difiere en el valor medio del que depende el peso independiente $\hat{w}_M^{b_0 \cdots b_i}$: los planos de discriminación son también paralelos y el entrenamiento será igual al del caso lineal.

Los vectores de pesos de los filtros serán proporcionales a los vectores

$$\hat{\mathbf{w}}^{*b_0 \cdots b_i} = \frac{\mathbf{w}^{*b_0 \cdots b_i}}{w^{*b_0 \cdots b_i}_0} = \left[1 \quad \frac{w^*_1}{w^{*b_0 \cdots b_i}_0} \quad \frac{w^*_{M-1}}{w^{*b_0 \cdots b_i}_0} \quad \cdots \quad h_0 m_{F\left(x^{b_0 \cdots b_i}\right)} \right]^T \qquad (4\text{-}33)$$

Es decir, el entrenamiento para el caso no lineal sin memoria no difiere del entrenamiento para el caso lineal.

A4.2.5 Entrenamiento en el caso de canal con no linealidad con memoria

En el caso de tener la no linealidad en la salida (no linealidad con memoria), el entrenamiento visto anteriormente no es válido; sin embargo, el algoritmo en escalera está diseñado para mejorar las prestaciones de los

---

[18] Si se entrena el peso independiente $w_M^{i,j}$ mediante LMS entonces la ecuación de actualización es $w_M^{b_0 \cdots b_i}[n] = (1-\mu) w_M^{b_0 \cdots b_i}[n-1] + \mu\left(d - \sum_{k=0}^{M-1} r[n-k] w_k^{b_0 \cdots b_i} \right)$. El límite de esta ecuación cuando n tiende a infinito es $\lim_{n \to \infty} w_M^{b_0 \cdots b_i}[n] = w_0^{b_0 \cdots b_i} m_{x^{b_0 \cdots b_i}}$

algoritmos basados en filtro FIR en este caso precisamente, y no en los anteriores[19].

Para el estudio que sigue se utiliza la expansión de Volterra de la señal de entrada al canal conservando todos sus términos en todo el desarrollo. Este procedimiento no es sino un caso particular del método de la perturbación [Ochi, 1990].

Se tiene un canal no lineal cuya salida en ausencia de ruido es la siguiente:

$$r[n] = G\left(\sum_{k=0}^{N-1} h_k x[n-k]\right) = G\left(\begin{bmatrix} x[n] \\ \vdots \\ x[n-N+1] \end{bmatrix}^T \mathbf{h}\right) = G\left(\mathbf{x}[n]^T \mathbf{h}\right) \qquad (4\text{-}34)$$

donde G(·) es una función no lineal del tipo expuesto en el capítulo segundo.

El dato $r[n]$ de entrada al receptor puede expandirse en una serie de Volterra en la siguiente forma:

$$\begin{aligned} r[n] = G\left(\mathbf{x}[n]^T \mathbf{h}\right) &= G(0) + \sum_{k_1=0}^{N-1} \frac{\partial G}{\partial x_k}\Big|_0 \, x_k + \dots \\ &+ \frac{1}{1!} \sum_{k_1=0}^{N-1} \dots \sum_{k_l=0}^{N-1} \frac{\partial^l G}{\partial x_{k_1} \cdots \partial x_{k_l}}\Big|_0 \, x_{k_1} \cdots x_{k_l} + \dots \end{aligned} \qquad (4\text{-}35)$$

siendo $\mathbf{0}=[0, \dots, 0]$ el origen de coordenadas del espacio R^N y $x_{k_i}=x[n-k_i]$, con $k_i=0,\dots N-1$.

Las derivadas parciales de G(·) respecto de x_{k_i} en la ecuación (4-35) toman la forma siguiente:

$$\begin{aligned} \frac{\partial^l G\left(\mathbf{x}^T\mathbf{h}\right)}{\partial x_{k_1} \cdots \partial x_{k_l}} &= \frac{\partial^{l-1}}{\partial x_{k_2} \cdots \partial x_{k_l}} \left(\frac{\partial \mathbf{x}^T\mathbf{h}}{\partial x_{k_1}} \frac{\partial G\left(\mathbf{x}^T\mathbf{h}\right)}{\partial \mathbf{x}^T\mathbf{h}} \right) = \\ &= \dots = h_{k_1} \cdots h_{k_l} G^{(l)}\left(\mathbf{x}^T\mathbf{h}\right) \end{aligned} \qquad (4\text{-}36)$$

$G^{(l)} = \frac{1}{(2l)!} \frac{d}{d^{2l}x} G(x)$ es la derivada de orden l evaluada en el origen de la función G(·), G(·) es impar, de manera que $\frac{\partial^{2n} G(x)}{\partial x^{2n}}\Big|_{x=0} = 0$ se puede escribir:

---

[19] Excepto porque en el algoritmo en escalera es posible insertar en forma directa la función de coste entrópica y porque el esquema en escalera disminuye el número de operaciones a realizar.

$$r[n] = G\left(\mathbf{x}[n]^T \mathbf{h}\right) = G^{(1)} \sum_{k_1=0}^{N-1} h_{k_1} x_k + \ldots$$
$$+ G^{(2\ell-1)} \sum_{k_1=0}^{N-1} \cdots \sum_{k_\ell=1}^{N-1} \left(h_{k_1} \cdots h_{k_\ell}\right)\left(x_{k_1} \cdots x_{k_\ell}\right) \tag{4-37}$$

La notación se puede simplificar haciendo uso del producto de Kroneker [Nowak, 1996] (véase el anexo 4.3).

Se define el vector $A^{(l)}$ de forma recursiva como

$$A^{(l)} = A^{(l-1)} \otimes A \tag{4-38}$$

siendo $\otimes$ el producto de Kronecker [Graham, 1981] y A una matriz cualquiera. Con esta notación los elementos de la derecha de la ecuación (4-37), que representan los términos de orden 2l-1 de la serie de Volterra, pueden reescribirse:

$$r[n] = G\left(\mathbf{x}[n]^T \mathbf{h}\right) = \sum_{l=1}^{\infty} G^{(2l-1)} \cdot \left(\mathbf{x}^{(2l-1)}\right)^T \mathbf{h}^{(2l-1)} \tag{4-39}$$

Una de las propiedades del producto de Kronecker (4.3) es que la transpuesta de la matriz producto es igual al producto de matrices transpuestas. Recursivamente se comprueba que $\left(\mathbf{x}^{(2l-1)}\right)^T = \left(\mathbf{x}^T\right)^{(2l-1)}$, con lo cual la expresión de $r_0[n]$ queda en la forma:

$$r[n] = \sum_{l=1}^{\infty} G^{(2l-1)} \cdot \left(\mathbf{x}^T\right)^{(2l-1)} \mathbf{h}^{(2l-1)} \tag{4-40}$$

El producto escalar en la ecuación (4-40) se puede simplificar de la siguiente manera:

$$\left(\mathbf{x}^T\right)^{(k)} \mathbf{h}^{(k)} = \left(\left(\mathbf{x}^T\right)^{(k-1)} \otimes \mathbf{x}^T\right)\left(\mathbf{h}^{(k-1)} \otimes \mathbf{h}\right)$$
$$= \left(\mathbf{x}^T\right)^{(k-1)} \mathbf{h}^{(k-1)} \otimes \mathbf{x}^T\mathbf{h} = \ldots = \left(\mathbf{x}^T\mathbf{h}\right)^k \tag{4-41}$$

En el último término el argumento del exponente de Kronecker es un escalar, con lo cual se convierte en un exponente ordinario. Con todo ello:

$$r[n] = \sum_{\ell=1}^{\infty} G^{(2l-1)}\left(\mathbf{x}^T\mathbf{h}\right)^{2l-1} \tag{4-42}$$

El vector de entrada al canal tiene, entonces, una expresión en forma de serie de Volterra:

$$\mathbf{r}[n]=\begin{bmatrix} r[n] \\ \vdots \\ r[n-M+1] \end{bmatrix}=\sum_{l=1}^{\infty} G^{(2l-1)}\begin{bmatrix} \left(\mathbf{x}^T[n]\mathbf{h}\right)^{2l-1} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\mathbf{h}\right)^{2l-1} \end{bmatrix} \qquad (4\text{-}43)$$

Se aplica esta entrada a un filtro FIR de M coeficientes:

$$o[n]=\mathbf{r}^T[n]\mathbf{w}=\begin{bmatrix} r[n] \\ \vdots \\ r[n-M+1] \end{bmatrix}^T \mathbf{w}=\sum_{l=1}^{\infty} G^{(2l-1)}\begin{bmatrix} \left(\mathbf{x}^T[n]\mathbf{h}\right)^{2l-1} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\mathbf{h}\right)^{2l-1} \end{bmatrix}^T \mathbf{w} \qquad (4\text{-}44)$$

Si la entrada al filtro de coeficientes $\mathbf{w}^{*\,b_0\ldots b_j}$ $b_0\ldots b_j$ está formada por vectores de la clase $b_0\ldots b_j$ (vectores con la forma de la ecuación (4-43) cuyo término x[n] es de la clase $b_0\ldots b_j$), entonces el cálculo de $\mathbf{R}^{b_0\ldots b_j}$ es el siguiente:

$$\mathbf{R}^{b_0\ldots b_j}=E\left\{\mathbf{r}^{b_0\ldots b_j}[n]\mathbf{r}^{b_0\ldots b_j\,T}[n]\right\}=$$

$$=E\left\{\begin{bmatrix} r^{b_0\ldots b_j}[n] \\ \vdots \\ r[n-M+1] \\ 1 \end{bmatrix}\begin{bmatrix} r^{b_0\ldots b_j}[n] \\ \vdots \\ r[n-M+1] \\ 1 \end{bmatrix}^T\right\}$$

$$=E\left(\sum_{l=1}^{\infty} G^{(2l-1)}\begin{bmatrix} \left(\mathbf{x}^{b_0\ldots b_j\,T}\mathbf{h}\right)^{2l-1} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\mathbf{h}\right)^{2l-1} \\ \left(G^{(2l-1)}\right)^{-1} \end{bmatrix}\sum_{m=1}^{\infty} G^{(2m-1)}\begin{bmatrix} \left(\mathbf{x}^{b_0\ldots b_j\,T}\mathbf{h}\right)^{2m-1} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\mathbf{h}\right)^{2m-1} \\ \left(G^{(2m-1)}\right)^{-1} \end{bmatrix}^T\right) \qquad (4\text{-}45)$$

$$=\sum_{l=1,m=1}^{\infty} G^{(2l-1)}G^{(2m-1)}E\left(\begin{bmatrix} \left(\mathbf{x}^{b_0\ldots b_j\,T}\mathbf{h}\right)^{2l-1} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\mathbf{h}\right)^{2l-1} \\ \left(G^{(2l-1)}\right)^{-1} \end{bmatrix}\begin{bmatrix} \left(\mathbf{x}^{b_0\ldots b_j\,T}\mathbf{h}\right)^{2m-1} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\mathbf{h}\right)^{2m-1} \\ \left(G^{(2m-1)}\right)^{-1} \end{bmatrix}^T\right)$$

Esta matriz tiene MxM elementos $\rho^{b_0\ldots b_j}_{2l-1,\,2m-1}$ cuya expresión es:

$$\rho_{l,m}^{b_0 \ldots b_j}[p,q] = E\left\{\left(\mathbf{h}^T\mathbf{x}[n-p]\right)^{2l-1}\left(\mathbf{x}^T[n-q]\mathbf{h}\right)^{2m-1}\right\}, \quad 0 > p,q >$$

$$\rho_{l,m}^{b_0 \ldots b_j}[p,0] = E\left\{\left(\mathbf{h}^T\mathbf{x}^{b_0 \ldots b_j}\right)^{2l-1}\left(\mathbf{x}^T[n-q]\mathbf{h}\right)^{2m-1}\right\}, \quad 0 > p > M$$

$$\rho_{l,m}^{b_0 \ldots b_j}[p,M] = E\left\{\left(\mathbf{x}^T[n-q]\mathbf{h}\right)^{2l-1}\right\} = 0 \quad 0 > p > M \tag{4-46}$$

$$\rho_{l,m}^{b_0 \ldots b_j}[0,0] = E\left\{\left(\mathbf{x}^{b_0 \ldots b_j\,T}\mathbf{h}\right)^{2(l+m-1)}\right\}$$

$$\rho_{1,1}^{b_0 \ldots b_j}[M,M] = 1$$

Con esto se puede escribir

$$\mathbf{R}^{b_0 \ldots b_j} = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} G^{(2l-1)}G^{(2m-1)}\mathbf{R}_{2l-1,2m-1}^{b_0 \ldots b_j} \tag{4-47}$$

siendo $\mathbf{R}_{2l-1,2m-1}^{b_0 \ldots b_j}$ la matriz de momentos cruzados de orden 2l-1,2m-1.

Los elementos de la matriz de autocorrelación (4-47) tienen la forma

$$\rho^{b_0 \ldots b_j}[p,q] = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} G^{(2l-1)}G^{(2m-1)}\rho_{2l-1,2m-1}^{b_0 \ldots b_j}[p,q] \tag{4-48}$$

Por otro lado, el cálculo de **P** es:

$$\mathbf{P} = E\left\{d^{b_0 \ldots b_j}\mathbf{r}^{b_0 \ldots b_j}[n]\right\} = E\left\{d^{b_0 \ldots b_j}\sum_{l=1}^{\infty}G^{(2l-1)}\begin{bmatrix}\left(\mathbf{x}^{b_0 \ldots b_j\,T}\right)^{(2l-1)}\mathbf{h}^{(2l-1)} \\ \vdots \\ \left(\mathbf{x}^T[n-M+1]\right)^{(2l-1)}\mathbf{h}^{(2l-1)}\end{bmatrix}\right.$$

$$\left. = \sum_{l=1}^{\infty}G^{(2l-1)}E\left\{\begin{bmatrix} d^{b_0 \ldots b_j}\left(\mathbf{x}^{b_0 \ldots b_j\,T}\right)^{(2l-1)}\mathbf{h}^{(2l-1)} \\ \vdots \\ d^{b_0 \ldots b_j}\left(\mathbf{x}^T[n-M+1]\right)^{(2l-1)}\mathbf{h}^{(2l-1)}\end{bmatrix}\right\} = \begin{bmatrix} p_{2l-1}^{b_0 \ldots b_j} \\ \vdots \\ \mathbf{0}\end{bmatrix}\right. \tag{4-49}$$

Con esto se puede escribir la ecuación del filtro transversal óptimo de Wiener en los siguientes términos:

$$\sum_{l=1}^{\infty}\sum_{m=1}^{\infty} G^{(2l-1)} G^{(2m-1)} \mathbf{R}_{2l-1}^{b_0 \dots b_j} \mathbf{w}^{b_0 \dots b_j} = \begin{bmatrix} \sum_{l=1}^{\infty} G^{(2l-1)} p_{2l-1}^{b_0 \dots b_j} \\ \vdots \\ 0 \end{bmatrix} \qquad (4\text{-}50)$$

que es la solución de Wiener para el canal no lineal con memoria.

Además, es fácil comprobar que

$$w_M^{b_0 \dots b_j} = E\left\{ G\left(x^{b_0 \dots b_j\, T}[n]\mathbf{h} \right) \right\} \qquad (4\text{-}51)$$

Se puede afirmar que:

1. Los hiperplanos de separación no son paralelos, aunque sus diferencias son muy pequeñas

Pasamos a examinar la semejanza de la solución para el filtro central con las soluciones para los demás filtros. Los elementos 0 > p,q >M de la matriz de autocorrelación no dependen de la clase de entrada al filtro. Pero los elementos p,q = 0 sí. En el caso lineal o con linealidad sin memoria, la primera columna de la matriz de autocorrelación es igual para todos los filtros, lo que significa que las soluciones a la ecuación matricial son paralelas. En este caso no es así debido a que los elementos de orden superior en la columna 0 de la matriz:

$$\rho_{l,m}^{b_0 \dots b_j}[p,0] = E\left\{ \left(\mathbf{h}^T \mathbf{x}^{b_0 \dots b_j} \right)^{2l-1} \left(\mathbf{x}^T[n-q]\mathbf{h} \right)^{2m-1} \right\}, \qquad 0 > p > M$$

son no nulos y dependen de potencias de $\mathbf{x}^{b_0 \dots b_j}$.

Todos los elementos de la última columna de la matriz $\mathbf{R}^{b_0 \dots b_j}$ son nulos excepto el primero y el último. Sustituyendo la solución anterior en la ecuación matricial resulta una ecuación matricial de orden M-1xM-1 que genera una recta que pasa por el origen. Se trata de saber cuáles son las diferencias entre los vectores directores de las rectas correspondientes a cada uno de los filtros. Esta diferencia viene exclusivamente determinada por los elementos $\rho_{l,m}^{b_0 \dots b_j}[p,]$ de la matriz de autocorrelación, que son los únicos que difieren de un filtro a otro: todos los vectores característicos de los hiperplanos 2 a M-1 en las ecuaciones $\mathbf{R}^{b_0 \dots b_j} \mathbf{w}^{*b_0 \dots b_j} = \mathbf{p}^{b_0 \dots b_j}$ son proporcionales excepto por su primer coeficiente. Pero la diferencia entre los vectores que generan la solución del primer filtro y los vectores que generan las demás soluciones está en los términos l>1 y m>1 de la serie. Es decir,

$$\rho_{1,m}^{b_0 \dots b_j}[p,0] - \rho_{1,m}[p,0] = \\ = \sum_{l=2}^{\infty}\sum_{m=2}^{\infty} G^{(2l-1)} G^{(2m-1)} \left(\rho_{2l-1,2m-1}^{b_0 \dots b_j}[p,0] - \rho_{2l-1,2m-1}[p,0]\right) \quad (4\text{-}52)$$

Las soluciones, por tanto, son iguales si los coeficientes de Volterra de orden mayor que uno son nulos (canal lineal). Su diferencia es pequeña siempre y cuando los coeficientes de Volterra sean pequeños[20], con independencia de la naturaleza de la no linealidad.

Para la no linealidad utilizada en las simulaciones:

$$G(x) = \frac{1}{\tanh\left(\frac{1}{10}\right)} \tanh\left(\frac{x}{10}\right)$$

que genera una compresión en torno al 10% la secuencia de coeficientes de Volterra es el de la Figura 4-1.

Para este caso se pueden despreciar los coeficientes de orden superior a tres, cuyo orden de magnitud es 10^{-5}. La diferencia entre los coeficientes de la matriz de autocorrelación del filtro central y la del filtro i,j, con un error absoluto de ese orden de magnitud es

2. La posición de las superficies sí depende de si el entrenamiento es o no discriminativo

Y además, el entrenamiento discriminativo conduce al sistema a una solución con menor riesgo en términos de tasa de fallos.

Si se opta por un entrenamiento discriminativo, la posición de la superficie de separación sólo está determinada por la estadística de los datos adyacentes: en las filas 2 a M, columna primera, de la matriz de autocorrelación sólo aparecen los estadísticos de los datos ±1. Sin embargo, si el entrenamiento no es discriminativo, las filas 2 a M, columna primera, contienen los estadísticos de todos los datos. Los vectores fila de la matriz aumentarán su ángulo de inclinación con respecto de la primera coordenada. Estos vectores están contenidos dentro de la superficie de separación: por tanto, la superficie de separación aumentará su ángulo respecto de la primera coordenada.

[20] Siempre y cuando los coeficientes h_j de la parte lineal del canal sean menores a la unidad.

La ecuación (4-52) es fácil de calcular si se conoce el valor cuadrático medio de los símbolos $x^{b_0 \cdots b_j}$. Siguiendo la notación, para el ejemplo 8-PAM utilizado se tienen los filtros $\mathbf{w}$, $\mathbf{w}^{b_0}$, $\mathbf{w}^{b_0 b_1}$. El filtro más alejado de $\mathbf{w}$ por la parte positiva es $\mathbf{w}^{11}$, seguido de $\mathbf{w}^{1}$ y $\mathbf{w}^{10}$, que es el más cercano (véase gráfico).

Las distancias entre elementos $\rho^{b_0 \cdots b_j}[p,0]$ para entrenamiento no discriminativo (divididas por el término $\rho^{b_0 \cdots b_j}[q,q] \approx \sigma_x^2$), suponiendo una interferencia intersimbólica del 20% y que los términos $\rho^{b_0 \cdots b_j}[p,q]$, $p \neq q$ son

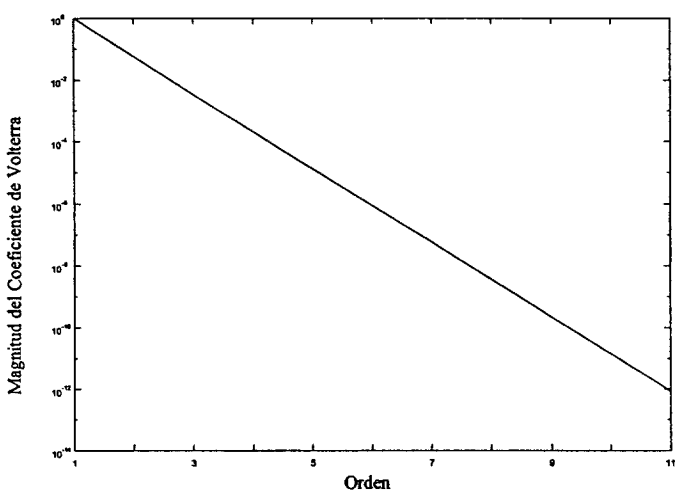


Figura 4-1. Coeficientes de volterra para la función $G(x) = \tanh\left(\frac{x}{10}\right) \Big/ \tanh\left(\frac{1}{10}\right)$

despreciables, resultan, para entrenamiento no discriminativo

$$\frac{\rho^{11}[p,0] - \rho\ [p,0]}{\rho^{11}[p,0]} \approx 3\%$$

$$\frac{\rho^{1}[p,0]-\rho\ [p,0]}{\rho^{01}[p,0]}=0$$

$$\frac{\rho^{01}[p,0]-\rho\ [p,0]}{\rho^{01}[q,q]}\approx 3\%$$

Nótese que para el filtro $\mathbf{w}^1$ la distancia es idénticamente nula: en efecto, las matrices son iguales en sus filas 2 a M: los datos a la entrada del filtro tienen la misma varianza y lo único que varía es su media, que no aparece en esas filas.

Las distancias para entrenamiento discriminativo son

$$\frac{\rho^{11}[p,0]-\rho\ [p,0]}{\rho\ [p,0]}\approx 2\%$$

$$\frac{\rho^{1}[p,0]-\rho\ [p,0]}{\rho\ [p,0]}\approx 4\%$$

$$\frac{\rho^{01}[p,0]-\rho\ [p,0]}{\rho\ [p,0]}\approx 10\%$$

los ángulos entre ecuaciones son 1.6° 0° y 1.6° para entrenamiento no discriminativo, y 1.5°, 3° y 6° para discriminativo. Por supuesto, el término en el

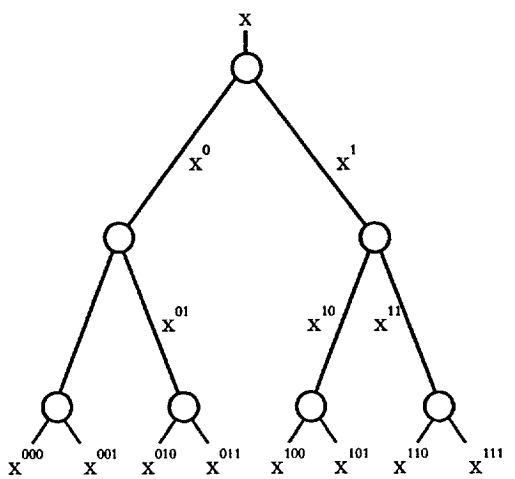


Figura 4-2. Distribución de los datos en el algoritmo en escalera.

denominador de los anteriores cocientes es mayor al que se utiliza y además, los otros términos, aunque pequeños, no son nulos, lo que significa que los ángulos reales serán menores.

Lo que nos interesa es la diferencia mayor. Esta diferencia no depende del número de niveles del alfabeto, sino sólo del nivel de saturación.

Si se opta por un entrenamiento discriminativo, debe hacerse de la siguiente manera: se calcula el gradiente del error para todos los filtros. La actualización de cada filtro $b_0...b_j$ es la combinación lineal del gradiente del error a la salida del filtro central más el gradiente a la salida del filtro $b_0...b_j$ que se pretende actualizar. La ponderación en cada filtro depende de la distancia entre la solución óptima para ese filtro y la del filtro central. Así, para el más alejado de los filtros se utiliza un 10% de su gradiente y un 90% del gradiente del filtro central. Para el más cercano se utilizará una proporción 1%-99%. La función de actualización de los pesos de cada uno de los filtros es:

$$\begin{aligned} \mathbf{w}_m^{b_0 \dots b_j}[n+1] &= \mathbf{w}_m^{b_0 \dots b_j}[n] + \eta\mu\Delta w_m^{b_0 \dots b_j} + (1-\eta)\mu\Delta w_m, \quad m < M \\ \mathbf{w}_M^{b_0 \dots b_j}[n+1] &= \mathbf{w}_M^{b_0 \dots b_j}[n] + \mu\Delta w_M^{b_0 \dots b_j} \end{aligned} \tag{4-53}$$

con $0.001 > \eta > 0.1$ Primero, el clasificador central lleva a los demás a una posición cercana a la óptima. Cuando el filtro del clasificador central haya convergido sólo actuará el término propio de cada filtro en la actualización y se ajustará la diferencia entre la solución de cada filtro y el central.

Cuando se utilice un entrenamiento no discriminativo debe tenerse en cuenta que los filtros de la segunda capa tendrán datos con la misma varianza que el central. Entonces debe procederse de la misma forma, pero aplicando a los filtros de la segunda capa la misma actualización que al filtro central. Además, la posición de los filtros que estén muy cerca del central no tiene porqué ser la más parecida, como se ve en el ejemplo 8-PAM. Sin embargo, la solución más recomendable es el entrenamiento discriminativo, porque el clasificador central tiene una posición determinada sólo por los datos adyacentes.

A4.2.6. Consideración final

El anterior análisis, aunque válido sólo para clasificadores basados en funciones lineales, sugiere la posibilidad de aplicar la misma técnica utilizando igualadores no lineales dispuestos en escalera tal como hemos visto. El clasificador central deberá llevar a los demás a una posición cercana a la de mínimo coste, y el ajuste fino se realizará independientemente para cada clasificador. Hay tres tipos de funciones candidatas a esta aproximación: el perceptrón multicapa, el filtro de Volterra y el GCMAC. El perceptrón multicapa, sin embargo, tiene una convergencia muy lenta y por ello el entrenamiento discriminativo puede ser perjudicial. Pero las técnicas para acelerar la

convergencia y evitar mínimos locales deben ser aplicadas al perceptrón central, y éste acelera a su vez a los demás.

Lo anterior no deja de ser una suposición, sobre todo en lo que respecta al perceptrón multicapa, porque no es posible hacer un análisis de su convergencia en la forma vista para el filtro FIR. De hecho, ya no somos capaces de repetir este análisis para otros objetivos como la entropía relativa. En el capítulo 5 se aplica esta técnica con filtros de Volterra aplicados en escalera. Por otra parte, en [Martínez/2, 1999] se presenta la aplicación del algoritmo en escalera utilizando funciones GCMAC [González, 1998].

Otros muchos esquemas no lineales no son susceptibles de ser utilizados en la misma forma (o por lo menos nosotros no somos capaces de verlo así). Entre éstos destacamos los basados en funciones de base radial y, en general, los sistemas basados en selección de muestras, los cuales, como se verá en el Capítulo 5, pueden ser guiados mediante otros sistemas.

ANEXO 4.3: PRODUCTO DE KRONECKER

A4.3.1 Producto de Kronecker

A matriz de orden m×n y B matriz de orden r×s. El producto de Kronecker se lleva a cabo de la siguiente manera:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix}$$

A4.3.2 Propiedades del producto de Kronecker

Linealidad

$$A \otimes (\alpha B) = \alpha(A \otimes B)$$

Demostración: el bloque (i,j) del producto $A \otimes (\alpha B)$ es

$$\begin{aligned} a_{ij}(\alpha B) &= \alpha[a_{ij}B] \\ &= \alpha[a_{ij}B] \\ &= \alpha[\text{bloque}\,(i,j)-\text{ésimo de}\, A \otimes B] \end{aligned}$$

Distributiva

$$\begin{aligned} (A+B)\otimes C &= A \otimes C + B \otimes C \\ A \otimes (B+C) &= A \otimes B + A \otimes C \end{aligned}$$

Demostración: el bloque (i,j) del producto $(A+B)\otimes C$ es

$$(a_{ij}+b_{ij})\otimes C \ (*)$$

El bloque (i,j) de $A \otimes C + B \otimes C$ es

$$(a_{ij}C + b_{ij}C) = (a_{ij}+b_{ij})C \ (**)$$

Como los bloques (*) y (**) son iguales para cada (i,j), queda demostrada la propiedad.

Asociativa

$$A \otimes (B \otimes C) = (A \otimes B) \otimes C$$

Demostración: inmediata. Las matrices tienen órdenes l×m, n×o, p×q. La matriz resultante tiene $l_{np} \times m_{oq}$ elementos de la forma $a_{gh}b_{ij}c_{kl}$.

Elemento neutro y elemento unidad

El elemento neutro es $0_{mn}=0_m \otimes 0_n$.
El elemento unidad es $I_{mn}=I_m \otimes I_n$.

Matriz producto de Kronecker transpuesta

$$(A \otimes B)^T = A^T \otimes B^T$$

Demostración: El bloque (i,j) de la matriz resultante es $a_{ij}B'$

Regla del producto mezclado

$(A \otimes B)(C \otimes D) = AC \otimes BD$
siempre que las dimensiones permitan el producto de matrices.

Demostración: El bloque (i,j) de la parte izquierda de la igualdad se obtiene tomando el producto de la i-ésima fila de bloques de A⊗B y la j-ésima columna de bloques de C⊗D, de la siguiente manera:

$$\begin{bmatrix} a_{i1}B & \cdots & a_{in}B \end{bmatrix} \begin{bmatrix} c_{1j}D \\ \vdots \\ c_{nj}D \end{bmatrix} =$$

$$= \sum_r a_{ir}c_{rj}BD$$

El bloque (i,j) de la parte derecha es (por definición del producto de Kronecker) $g_{ij}BD$ donde g_{ij} es el elemento (i,j) de la matriz AC, pero por la regla de la multiplicación de matrices

$$g_{ij} = \sum_r a_{ir}c_{rj} .$$

Como los bloques (i,j)-ésimos son iguales, se cumple la igualdad, c.q.d.

Inversa del producto

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

Demostración: Usando la regla del producto mezclado

$$(A \otimes B)\,(A \otimes B)^{-1} = AA^{-1} \otimes BB^{-1} = I_m \otimes I_n = I_{nm}$$

Vectores y valores propios de la matriz producto

Sea A una matriz con valores propios λ_i y vectores propios x_i y sea B una matriz con valores propios μ_j y vectores propios y_j. Entonces la matriz resultante de $(A \otimes B)$ tiene valores propios $\lambda_i \mu_j$ y vectores propios $x_i \otimes y_j$.

Demostración:

$$\begin{aligned}(A \otimes B)(x_i \otimes y_j) &= (Ax_i) \otimes (By_j) \\ &= (\lambda_i x_i) \otimes (\mu_j y_j) \\ &= (\lambda_i \mu_j)(x_i \otimes y_j)\end{aligned}$$

A4.3.3 Exponente de Kronecker

$$A^{(l)} = A^{(l-1)} \otimes A \text{ (recursivamente)}$$

Propiedades

Se deducen de forma inmediata de las anteriores.

$(A^{(l)})^T = (A^T)^{(l)}$
$A^{(l)} B^{(l)} = (AB)^{(l)}$
$A^{(l)} B^{(m)} = A^{(l-m-1)} \otimes (AB)^{(m)}$

5. IGUALADORES CON ELEMENTOS NO LINEALES

5.1. INTRODUCCIÓN

El uso de fronteras no lineales quedó justificado en el Capítulo 2, donde se muestran algunos ejemplos en los que se encuentra numéricamente la frontera de Bayes, con la que se consigue la mínima probabilidad de error. En canales de fase mínima, aún cuando es razonable el uso de igualadores lineales, éstos alcanzan probabilidades de error bastante mayores que los no lineales cuando la potencia de ruido es baja. Cuando la potencia de ruido es muy alta la frontera de Bayes es casi lineal. Dicho de otra manera, los sistemas lineales saturan sus prestaciones cuando la SNR aumenta.

Por otro lado, si el canal presenta no linealidades, sobre todo cuando estas no linealidades son con memoria, en algunas regiones del espacio de datos la distribución de los datos hace que los esquemas basados en funciones lineales tengan bajas prestaciones. Los esquemas no lineales, al presentar mejor aproximación a la frontera de Bayes, consiguen tasas de error más bajas.

En cambio, necesitan una potencia de cálculo mucho mayor y necesitan de muchos más datos para alcanzar la convergencia. Entre otras cosas, esto supone una menor capacidad de seguimiento de canales no estacionarios.

Para enfrentarse con esta dicotomía, en este capítulo se propone utilizar el esquema en escalera visto en los anteriores capítulos, pero sustituyendo las funciones lineales por otras no lineales. Con ello se obtiene alguna ventaja en

cuanto a seguimiento de canales no estacionarios.

En canales no lineales los datos que tienen menor amplitud pueden ser clasificados con esquemas lineales con una tasa de error menor al riesgo de Bayes de los datos que tengan amplitudes mayores: el uso de funciones no lineales puede que mejore la tasa de error en las zonas de baja amplitud, pero la mejora será despreciable si en las zonas de alta amplitud la tasa de fallos es, por ejemplo, de un orden de magnitud mayor.

En ese caso, deben utilizarse funciones lineales allí donde los datos presenten riesgo de Bayes bajo y funciones no lineales allí donde la distorsión no lineal haga que la tasa de error aumente.

El inconveniente de aplicar esta segunda opción es que el sistema no lineal no podrá ser guiado mediante otro sistema igual que esté en el centro de los datos. Sin embargo, la aproximación que llamamos Selección de Muestras, es un método guiado por una frontera de clasificación de referencia. En esta aproximación, esta frontera de referencia puede ser lineal y puede utilizar el esquema en escalera, lo que aumenta la capacidad de seguimiento del sistema.

5.2. ALGORITMOS COMPLETAMENTE NO LINEALES

El algoritmo en escalera completamente no lineal consiste en la inserción de funciones no lineales iguales en todos los clasificadores del algoritmo. El entrenamiento del clasificador central se traslada a los siguientes, entrenándose éstos en la forma vista en el Capítulo 4. Se presentan brevemente dos aproximaciones no lineales que siguen la estructura del esquema en escalera. La primera se basa en polinomios de Volterra y la segunda está basada en el esquema Generalyzed Cerebelar Model Arithmetic Computer [González, 1998].

5.2.1. Polinomios de Volterra

El esquema basado en polinomios de Volterra tiene el inconveniente de que la señal de entrada debe tener una amplitud normalizada, de forma que se pueda controlar el margen dinámico de la salida. En las simulaciones, la potencia de la entrada a cada uno de los clasificadores se normaliza utilizando la siguiente fórmula:

$$\hat{x} = \frac{x}{\bar{x}} - \frac{\bar{x}}{2} \tag{5-1}$$

siendo $\bar{x}$ el valor medio de los datos a la entrada del polinomio de Volterra.

Si el espacio de datos de entrada es de orden M y el filtro de Volterra de

cada uno de los clasificadores en escalera es de orden S, el número K total de pesos a actualizar para cada uno de los filtros es $K = \sum_{s=0}^{S} M^s$.

La expresión del polinomio de Volterra para el clasificador $b_0...b_j$ es [Nowak, 1996]:

$$o^{b_1 \dots b_j} = \sum_{l=1}^{L} \mathbf{w}(l)^{b_0 \dots b_j} \mathbf{x}^{(l)} + w_K \qquad (5\text{-}2)$$

$\mathbf{x}^{(l)}$ denota el exponente de Kronecker (véase Anexo 4.2): $\mathbf{x}^{(1)} = \mathbf{x}$ y, recursivamente, $\mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} \otimes \mathbf{x}$. De esta manera se consigue una expresión compacta y fácil de programar para la serie. $\mathbf{w}(l)$ es el vector de pesos del elemento de orden l del filtro de Volterra

Para hacer que el filtro de Volterra sea adaptativo frente a canales no estacionarios, se entrena mediante gradiente.

El inconveniente de este esquema es su gran complejidad. Por ejemplo, para M=2 y S=3, el número total de pesos es K=15. Para S=1 (filtro lineal), el número es K=M+1=3. Además, este tipo de filtros es difícil de entrenar y, para los canales no estacionarios que se simulan en esta tesis, la tasa de desenganche es muy alta. Para el canal que se utiliza en este capítulo, y que tiene una parte lineal con la expresión $h[n] = \delta[n] + h_1[n]\delta[n-1]$

con

$$h_1[n] = \begin{cases} 0.2 & 1 < n < 3000 \\ 0.995 h_1[n-1] - 0.001 & 3001 < n < 5000 \end{cases}$$

el esquema en escalera con filtros de Volterra de orden 5 se desengancha en 125 de 300 realizaciones. Por otro lado, la tasa de error no supera a la del filtro lineal.

Existe un gran número de referencias (ya citadas) que tratan de la estabilidad de los filtros de Volterra. En esta tesis, y vistas las dificultades que se han encontrado en la realización de esquemas en escalera mediante filtros de Volterra, se presentan métodos alternativos que pretenden mejorar las prestaciones de aquéllos.

5.2.2. Generalised Cerebellar Model Arithmetic Computer (GCMAC)

El GCMAC [González, 1997], es una estructura que posee propiedades de igualación no lineal a la vez que tiene una estructura lineal. Esta estructura necesita muy pocos cálculos por dato para llegar a la convergencia. El igualador de Volterra necesita 1000 veces más operaciones para llegar a la convergencia.

Además, los tiempos de convergencia de ambos igualadores son similares [González, 1998].

En el GCMAC, la entrada es cuantificada en L valores discretos. Cada muestra direcciona una memoria ROM; cada celda de esta memoria contiene un conjunto de punteros a una memoria RAM. Las posiciones de la memoria RAM deben ser diferentes para cada muestra cuantificada. El número de posiciones de la RAM direccionadas debe ser una constante $\rho>2$. Además, el grupo de celdas direccionadas por dos valores consecutivos de la entrada cuantificada deben diferir en una posición de memoria

El GCMAC aproxima la función no lineal deseada para la igualación mediante el uso de funciones de base local (LBF) solapadas entre ellas. Las LBF se sitúan en intervalos fijos que digitalizan el espacio de entrada en celdas. Las LBF se definen como regiones hiperparalelepípedas iguales, cuyo tamaño está determinado por el vector $\rho = [\rho_1, \cdots, \rho_p]^T$, donde $1<\rho_i<L_i$. Cuanto menor sea ρ_i, máyor el número de LBFs a lo largo del eje i. L_i es el número de niveles de la amplitud de la entrada i.

La relación entrada/salida del GCMAC puede ser descompuesta en dos mapeos consecutivos. Primero, se selecciona el conjunto de celdas apropiado según la entrada y el segundo lleva a cabo una combinación lineal del valor de éstas ($\phi_i(x)$) con un vector **w**, que produce la salida:

$$o = \mathbf{w}^T \phi(x) \tag{5-3}$$

La actualización de los pesos se hace mediante un método de gradiente, de manera que puede aplicarse a un esquema en escalera [Martínez, 1999].

Igual que en los esquemas en escalera lineales, el entrenamiento del igualador central puede aplicarse a los siguientes según la regla (4-53), pág. 121:

$$\begin{aligned} w_k^{b_0 \cdots b_i}[n+1] &= w_k^{b_0 \cdots b_i}[n] + \eta^{b_0 \cdots b_i} \mu \Delta w_k^{b_0 \cdots b_i} \\ &\quad + \left(1 - \eta^{b_0 \cdots b_i}\right) \mu \Delta w_k, \qquad k < M \\ w_M^{b_0 \cdots b_i}[n+1] &= w_M^{b_0 \cdots b_i}[n] + \mu \Delta w_M^{b_0 \cdots b_i} \end{aligned} \tag{5-4}$$

La (Figura) muestra la tasa de convergencia y la capacidad de seguimiento de un igualador en escalera basado en GCMAC. $\rho=(16, 16)$ y el número de niveles de la cuantificación es 2^6.

- Respuesta impulsional de la parte lineal del canal:

$$h[n] = \delta[n] + h_1[n]\delta[n-1]$$

con

$$h_1[n] = \begin{cases} 0.2 & 1 < n < 3000 \\ 0.995h_1[n-1] - 0.001 & 3001 < n < 5000 \end{cases}$$

- no linealidad $G(x) = \tanh(x/\xi)/\tanh(1/\xi)$ con $\xi = 7$.
- SNR= 30 dB

La señal emitida es 8-PAM con amplitudes {±1, ±3, ±5, ±7} ("back-off " de 0 dB).

El seguimiento es lento; sin embargo, éste se puede mejorar optimizando la combinación de entrenamientos. Además, se pueden arbitrar procedimientos que hagan adaptativo el valor del vector ρ, de manera que conceda más importancia a unas dimensiones y menos a otras.

Por otra parte, la tasa de fallos se acerca, en el estado estacionario, a 10^{-4}, mientras que las aproximaciones presentadas en esta tesis tienen una tasa de fallos cerca del doble.

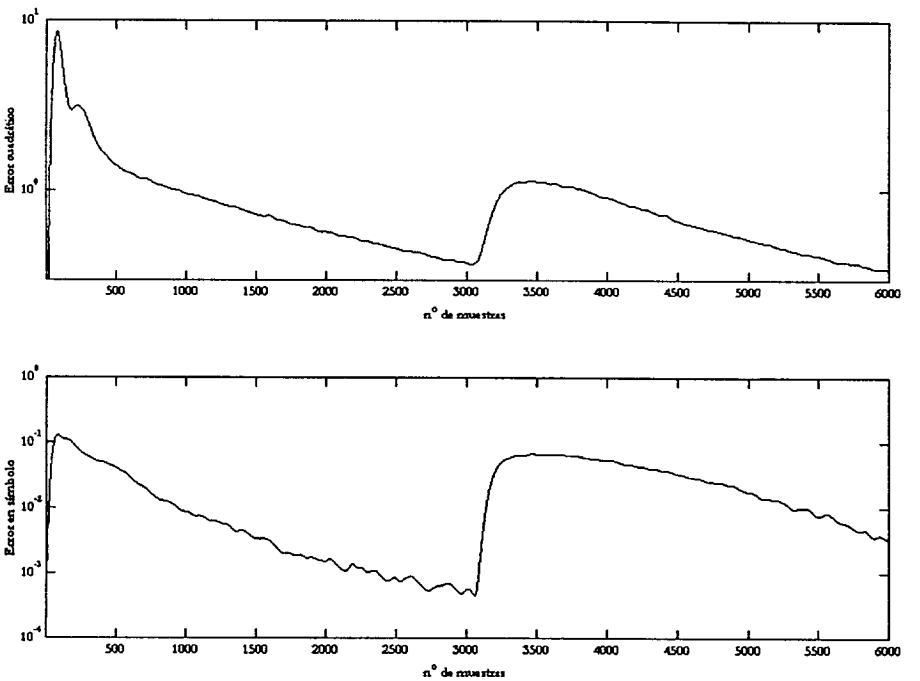


Figura 5-1. Error cuadrático y tasa de error de la salida del igualador en escalera basado en GCMAC..

5.3. ALGORITMOS PARCIALMENTE NO LINEALES

5.3.1. Polinomios de Volterra

En el apartado 5.2.1 se ha visto que el clasificador basado en polinomios de Volterra tiene un gran número de elementos. Sin embargo, es posible reducir algo este número si se utilizan polinomios ortogonales [Schetzen, 1981]. El filtro central y los filtros siguientes del esquema en escalera que reciban datos que puedan ser clasificados linealmente con una probabilidad de error baja deben construirse utilizando sólo el polinomio de orden 1.

Si los pesos del filtro se ordenan según se obtienen de la ecuación (5-2), el peso de orden cero tiene el subíndice K, y los de orden 1 los subíndices 1 a M (M es el número de dimensiones de los datos de entrada). Estos filtros se entrenan según la regla (4-53).

En los clasificadores en los que la probabilidad de error sea más alta, hay que colocar elementos de orden superior. Se puede combinar el entrenamiento de la misma forma que antes, sólo que los elementos de orden superior de un clasificador se entrenarán sólo con el gradiente calculado a partir de su propia salida, y no de la salida central.

Como los polinomios son ortogonales, el entrenamiento de un polinomio no afectará al de los otros.

Se puede escoger la serie de polinomios ortogonales según la norma

$$\left\langle P_i(x), P_j(x) \right\rangle = \int_{-1}^{1} g(x) P_i(x) P_j(x) dx \qquad (5\text{-}5)$$

con función de ponderación g(x) = 1. Se tiene entonces la serie de polinomios ortonormales de Legendre:

$$P_0(x) = \sqrt{\frac{1}{2}}$$

$$P_1(x) = \sqrt{\frac{3}{2}}x$$

$$P_2(x) = \frac{3}{2}\sqrt{\frac{5}{2}}\left(x^2 - \frac{1}{3}\right)$$

$$P_3(x) = \frac{5}{2}\sqrt{\frac{7}{2}}\left(x^3 - x\frac{3}{5}\right)$$

Esta serie se puede utilizar cuando los diferentes filtros de Volterra del esquema en escalera tengan órdenes crecientes; pero no produce ventajas en el caso en que se opte por filtros de orden igual.

Por otro lado, si hay simetría en los datos, únicamente son necesarios los polinomios de orden impar.

5.3.2. Redes de Funciones de Base Radial

El igualador basado en funciones de base radial (RBF) tiene el inconveniente de que no se conoce a priori el número adecuado de centroides a emplear. Si se emplean pocos, entonces se corre el riesgo de que la aproximación generalice mal, mientras que si se emplean demasiados, la red se adapta demasiado lentamente a la estadística de los datos.

Otro problema es la elección de la forma de las funciones. Si se escoge una red basada en funciones Gaussianas

$$o(x) = \sum_{k=0}^{K-1} w_k e^{-\frac{\|x-c\|^2}{2\sigma^2}} + w_K \qquad (5\text{-}6)$$

el parámetro de tamaño es σ. Un método es entrenar por gradiente este parámetro. Sin embargo, la velocidad de convergencia es muy baja, y, si se escogen desviaciones típicas diferentes para las M funciones del sumatorio, para aumentar las prestaciones, puede caerse en mínimos locales.

Recientemente se ha introducido el concepto de Centroide Crítico en [Sancho, 1999], [Lyhyaoui/1, 1999], [Lyhyaoui/2, 1999]. Esta aproximación resuelve eficientemente el problema de la elección del número de centroides. Para ello es necesario un periodo de entrenamiento de un número inicial de centroides que debe estimarse experimentalmente, y además, la cantidad de centroides escogida, que será válida para una determinada situación, será constante a lo largo del tiempo. Pero, con una ligera modificación, estas dos restricciones pueden

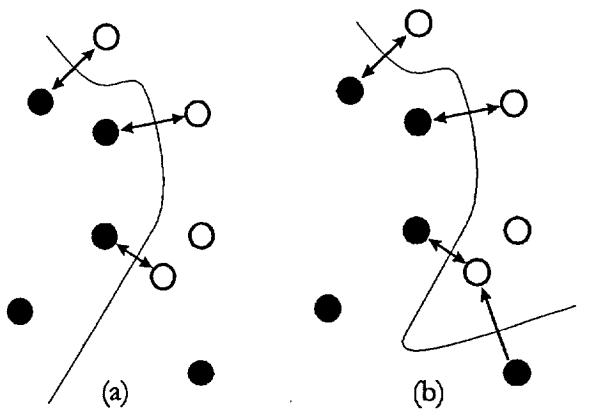


Figura 5-1. a) Para cada centroide se busca el más cercano de la clase contraria. Si la máxima proximidad se cumple en ambos sentidos (centroides unidos por flechas), el par de centroide es crítico. Se construye con ellos una frontera, que no clasifica correctamente, en este caso, a todos los centroides. b) Se encuentra el centroide mal clasificado más cercano a uno de la clase contraria y se incluye en el conjunto. Se verifica si el nuevo conjunto clasifica bien todos los centroides, en cuyo caso finaliza la selección. Si no, se repite b).

relajarse.

En cuanto a los valores de σ, una aproximación que produce buenos resultados es estimar sus valores en función de las distancias entre centroides de clase diferente, teniendo en cuenta la sensibilidad de la frontera de clasificación frente a cambios en los pesos de la capa de salida de la red.

5.3.2.a Centroides críticos

Sea un espacio de datos de dos clases diferentes. Supóngase un conjunto de centroides críticos entrenados mediante la cuantificación vectorial de datos descrita en 2.3.4.c. Se establece la clasificación de datos basada en el centroide más próximo: un punto en el espacio de datos pertenece a la clase del centroide del que está más cerca.

Se define el conjunto de centroides críticos como el conjunto de centroides de mínimo cardinal capaz de clasificar correctamente a todos los demás.

Esta definición es constructiva[21]: es decir, a partir de ella se puede hallar el conjunto de centroides críticos correspondiente a la cuantificación vectorial. Para ello:

- para cada centroide se encuentra el centroide más cercano de la clase contraria. Si dos centroides son recíprocamente el más cercano al otro, se

[21] Cuando se extiende este concepto al conjunto de muestras o datos, la definición ya no es constructiva. Más adelante se presenta el concepto de Muestra Crítica. Desgraciadamente, no disponemos de una definición que sea tan elegante y sencilla como ésta y que lleve implícita la construcción del conjunto.

considera que son críticos.

- Una vez hallados todos los pares críticos, se verifica si este conjunto es capaz de clasificar correctamente a los demás centroides mediante el método del centroide más próximo. Si es así, el conjunto queda determinado.
- En caso contrario, se introduce el centroide mal clasificado más próximo a un centroide crítico de clase contraria en el conjunto, y se repite el paso anterior, hasta que todos los centroides queden bien clasificados.

5.3.2.b Eliminación y división de centroides

Un método sencillo para evitar que haya demasiados o demasiado pocos centroides si cambia la estadística del canal una vez se ha llevado a cabo la selección de centroides críticos, es eliminarlos o dividirlos teniendo en cuenta su utilización.

El método de clusterización aplicado se inicializa con un número (dependiendo del número de dimensiones de los datos y del orden del canal) de centroides $\{c_i\}$ distribuidos aleatoriamente en el espacio. Se utiliza una distribución uniforme en el espacio que ocuparían los datos si el canal fuera ideal (lineal y sin interferencia intersimbólica). Se etiquetan los centroides de cada clase. Cada vez que se adquiere un dato x[n], la actualización de los centroides, dependiendo de si la clase del dato es diferente o igual a la del centroide) se lleva a cabo según la norma:

$$c_i[n+1] = \begin{cases} c_i[n] + q(c_i, x)\,(x[n] - c_i[n]), & \text{clases iguales} \\ c_i[n], & \text{clases distintas} \end{cases} \qquad (5\text{-}7)$$

siendo

$$q(c_i, c_G) = \mu \exp\left(\| c_i - c_G \|^2 / 2\sigma^2 \right) \qquad (5\text{-}8)$$

un término que produce mayor actualización cuanto más cerca esté el centroide c_i del centroide c_G más cercano al dato.

El nivel de actualización se mide acumulando la cantidad $U_i = \sum_n q(c_i, c_G)$ para cada centroide y dato adquirido.

Después de un número de actualizaciones, aquellos centroides que tienen U_i muy por debajo de la media (se toma como umbral el 10% de la media en las simulaciones) se eliminan.

Si un centroide tiene U_i mayor que el doble de la media, se divide, introduciendo en sus posiciones un término aleatorio para separar los nuevos

centroides, y se reparte U_i entre ambos. Esta sencilla técnica previene el uso de demasiados centroides y acelera la clusterización, a la vez que evita que el espacio se quede con demasiado pocos centroides si la estadística de los datos cambia.

Por otro lado, se puede sustituir la función (5-8) por una aproximación lineal, lo que no afecta a las prestaciones, y disminuye el coste computacional.

5.3.2.c Estimación de σ_i.

Un criterio de elección de σ_i es hacerla proporcional a la distancia del centroide correspondiente a la frontera de clasificación, de manera que ésta esté determinada principalmente por su función de base radial.

La distancia de la frontera al centroide será aproximadamente igual a la mitad de la distancia del centroide al más próximo de la clase contraria. Hay que tener en cuenta que a causa de que se escogen no sólo pares críticos, sino también centroides complementarios para formar el conjunto de centroides críticos, puede que el centroide c_i para el que estamos calculando σ_i sea el más cercano a más de un centroide. En ese caso, hay que escoger, para este subconjunto de centroides, la distancia mínima, a efectos del cálculo de sus σ. En caso contrario, la frontera no estará delimitada correctamente. La Figura 5-2 esquematiza el problema y su corrección. Se tiene un centroide c_i de una de las dos clases, cuyo centroide más cercano es c_j. Se escoge para σ_i y σ_j un valor proporcional a su semidistancia. Si se desprecia el efecto de las otras "gaussianas", la frontera se sitúa en el punto medio de los dos centroides. Pero el centroide c_k tiene como centroide más cercano de la clase contraria c_i. Se escoge para c_k la semidistancia, pero como para c_i ya se ha escogido otro valor menor, la frontera queda mal delimitada. Para que la frontera entre c_i y c_k esté en el punto medio de los dos centroides, se escoge para c_k $\sigma_k = \sigma_i$.

Lo importante ahora es escoger una proporción adecuada para σ_i. A tal efecto analizamos primero el caso correspondiente a una red de solo dos centroides en un espacio unidimensional. Para esta red, la salida será:

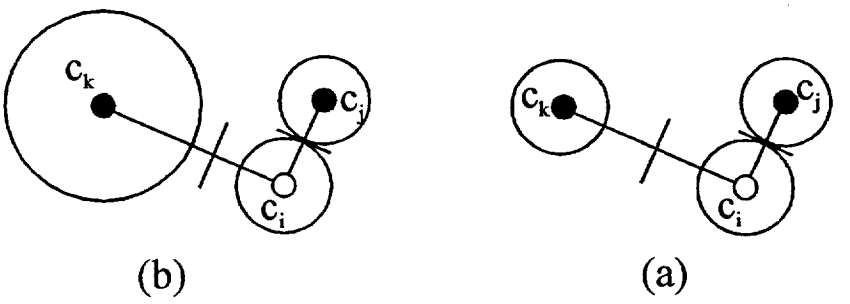


Figura 5-2. a) Inconveniente de usar la distancia entre centroides para estimar la σ; b) corrección de la frontera usando la distancia mínima.

$$o = w_1 e^{-\frac{\|x-c_1\|^2}{2(\delta d)^2}} + w_2 e^{-\frac{\|x-c_2\|^2}{2(\delta d)^2}} \tag{5-9}$$

Si se sitúa el origen de coordenadas en el punto medio de los dos centroides se tiene $c_1 = -c_2 = d$. La solución que proporciona la cuantificación vectorial es $w_1 = -w_2 = 1$, con lo que la frontera estará situada en x=0. Se acepta como válida esta solución a efectos del siguiente análisis.

La sensibilidad de la salida con respecto a la posición de la frontera es

$$s = \frac{\partial o}{\partial x} = -\frac{x-d}{\delta^2 d^2} e^{-\frac{\|x-d\|^2}{2\delta^2 d^2}} + \frac{x+d}{\delta^2 d^2} e^{-\frac{\|x+d\|^2}{2\delta^2 d^2}}$$

$$\frac{\partial o}{\partial x}(x=0) = \frac{e^{-\frac{1}{2\alpha^2}}}{\delta^2 d} \tag{5-10}$$

Un criterio razonable de elección del valor de δ es minimizar la sensibilidad de la frontera [Lyhyaoui/1, 1999]. Se escoge, pues, de manera que se cumpla

$$\frac{\partial s}{\partial \delta} = 2\left(\frac{1}{\delta^2} - 2\right)\frac{e^{-\frac{1}{2\delta^2}}}{\delta^3 d} = 0$$

De aquí se obtiene $\delta_0 = \sqrt{2/2}$.

La interpretación de este resultado es la siguiente. La pendiente de o(x) depende del valor de δ (véase la Figura 5-3.a). Para valores menores a $\sqrt{2}/2$ la pendiente en torno a la frontera será muy pequeña. Pequeñas variaciones de w_1 o w_2 harán que la frontera cambie bruscamente su posición. Para $\delta_0 = \sqrt{2/2}$ la pendiente es máxima. Las variaciones de los pesos guardan una relación aproximadamente lineal con la variación de la posición de la frontera. Para valores mayores, la pendiente de la salida disminuye lentamente (véase Figura 5-3.b).

Este análisis es extensivo a un espacio multidimensional. Primero, supóngase que la frontera en la vecindad de los dos centroides c_i y c_j de la ecuación solamente depende de éstos. Por otro lado, situar el origen de coordenadas de un sistema de K dimensiones de manera que las dos gaussianas estén situadas sobre el eje x_k en las coordenadas $\pm d$ no supone pérdida de generalidad.

La aproximación a la salida del clasificador es

$$o = w_i e^{-\frac{\|x-c_i\|^2}{2(\delta d)^2}} + w_j e^{-\frac{\|x-c_j\|^2}{2(\delta d)^2}} \tag{5-11}$$

Los puntos del plano perpendicular al eje x_k que pasa por el origen (donde se sitúa localmente la frontera) cumplen:

$$\left\|x\right|_{x_k=0} - c_i\| = \left\|x\right|_{x_k=0} - c_j\| = \left\|x\right|_{x_k=0}\| + d^2 = k_x^2 + d^2 \tag{5-12}$$

k_x es la distancia desde el punto medio de los centroides hasta un punto del plano perpendicular al eje x_k.

Se quiere ahora minimizar la sensibilidad de la frontera en este punto. Si se deriva (5-11) respecto de δ se obtiene:

$$s = \frac{\partial o}{\partial x}(x_k = 0) = \frac{1}{\delta^2 d} e^{-\frac{|k_x/d|^2+1}{2\delta^2}}$$

y de aquí, haciendo $\partial s/\partial\delta = 0$ se obtiene

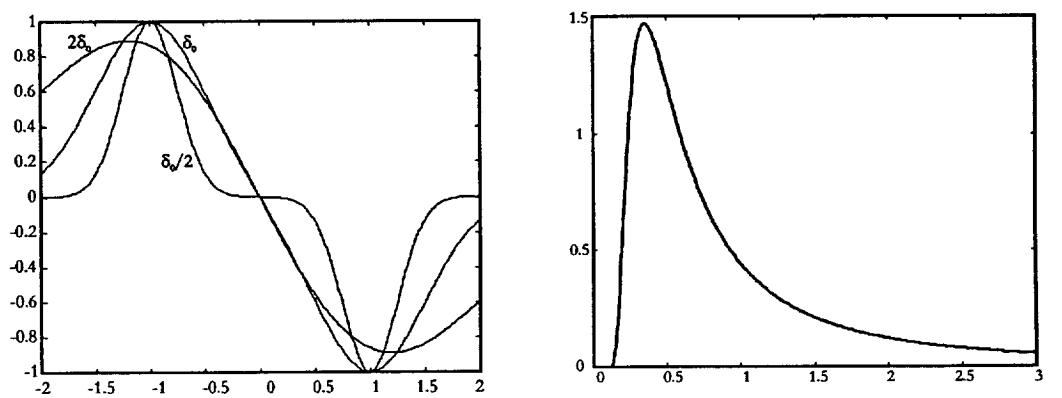


Figura 5-3. Salida de la red RBF para diferentes valores de δ (izquierda) y derivada de o en el origen para diferentes valores de δ.

$$\delta_{0x} = \frac{\sqrt{2}}{2}\sqrt{\frac{k_x^2 + d^2}{d^2}} \tag{5-13}$$

La siguiente cuestión es cómo escoger la k_x adecuada. Si se puede suponer que los datos en torno a los centroides están distribuidos según una densidad de probabilidad gaussiana, entonces la desviación típica de éstos está en torno a d. Si se escoge $\sqrt{k_x^2 + d^2} = 2d$, la zona de mínima sensibilidad está a una distancia en torno a dos veces la desviación típica de los datos: más del 90% de los datos estarán dentro de la zona de la frontera que tiene una sensibilidad pequeña. Para este caso, $\delta_{0x} = \sqrt{2}$.

5.3.2.d Red RBF de centroides críticos

A efectos de comparación, se simula una red de RBF. Esta red se inserta en el esquema en escalera como el clasificador del par de símbolos más exteriores de la constelación {±5 y ±7}. Se aplica a todas las RBF la estimación de su desviación típica descrita arriba, con el valor teórico de $\delta_{0x} = \sqrt{2}$ obtenido. Se hacen simulaciones para 15+15 centroides. La fase guiada dura 250 muestras y después se conmuta a funcionamiento dirigido por decisión.

Después de un periodo inicial de 1000 muestras se lleva a cabo una selección de centroides críticos. Se entrena la red resultante mediante gradiente.

La simulación se representa en la Figura 5-15, la Figura 5-16 y la Figura 5-17, en comparación con las simulaciones de la red de Muestras Críticas, que se expone a continuación.

5.4. SELECCIÓN DE MUESTRAS

5.4.1. Introducción: Máquinas de Vectores de Soporte

El concepto de Máquina de Vectores de Soporte (SVM) es introducido por Vladimir Naumovich Vapnik hacia 1979 [Vapnik, 1979]; sin embargo, hasta mediados de la década de los 90 no se le presta una especial atención a este concepto [Vapnik, 1995].

Las máquinas clásicas de clasificación se construyen de la siguiente manera: se construye una familia de funciones que describan una frontera (lineal o no) de separación entre clases; se entrena esta frontera utilizando para ello un algoritmo de aprendizaje (basado en estimación de gradiente o en otras técnicas). La solución a la que se llega se obtiene utilizando información de todos y cada

uno de los vectores del conjunto de entrenamiento utilizado.

El enfoque de las SVM es diferente. Desde un punto de vista cualitativo se puede afirmar que este tipo de máquinas se construyen de la siguiente manera: se toma un conjunto de vectores de entrenamiento y, mediante algún procedimiento, se seleccionan aquellos vectores de ambas clases que resulten más propicios para conformar una frontera de clasificación. Estos son los llamados Vectores de Soporte (SV). Entonces se construye una máquina de clasificación utilizando estos vectores.

En la práctica, se puede utilizar una combinación de ambos sistemas: primero se hallan los vectores de soporte; se construye con ellos una máquina de clasificación y luego se entrena mediante un algoritmo de gradiente. Por ejemplo, en [Scholkopf, 1997] se construye una máquina de clasificación con RBFs cuyos centros son los vectores de soporte encontrados previamente. Después se entrena este conjunto de RBFs utilizando retropropagación.

Para una máquina de clasificación basada en la función lineal

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{5-14}$$

el hiperplano óptimo de separación es aquél que verifica

$$\min\|\mathbf{w}\| \tag{5-15}$$

con las restricciones

$$d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 \tag{5-16}$$

siendo $\mathbf{x}_i$ la i-ésima muestra, y d_i su clase.

La solución de este problema la proporciona la función lagrangiana:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N}\alpha_i\left\{\left[\mathbf{w}_i^T\mathbf{x} + b\right]d_i - 1\right\} \tag{5-17}$$

donde α_i son los multiplicadores de Lagrange. La función minimiza la norma del vector $\mathbf{w}$ del hiperplano y maximiza la distancia entre el hiperplano y los datos.

Por otro lado, es fácil ver que si un dato verifica $d_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) = 1$ su distancia al plano es $d=1/\mathbf{w}$. Es decir, la ecuación (5-17) maximiza la distancia de aquellas muestras que están más cerca del plano.

Se puede demostrar que el hiperplano óptimo es una combinación lineal de las muestras de la forma

$$\mathbf{w}(\alpha) = \sum_{i=1}^{N}\alpha_i d_i x_i \tag{5-18}$$

y por lo tanto se puede escribir

$$W(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j \tag{5-19}$$

con la restricción

$$\alpha_i > 0 \tag{5-20}$$

Si se resuelve este problema de minimización cuadrática, los multiplicadores de Lagrange de los datos que verifiquen $y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) > 1$ serán nulos, en tanto que los multiplicadores de los datos que verifiquen la igualdad, no. El hiperplano óptimo es una combinación lineal de éstos.

La anterior solución está restringida a un problema linealmente separable. Si no existe separabilidad lineal, se generaliza el problema introduciendo unos términos ξ_i de error o pérdida en la ecuación (5-16):

$$d_i\left(x_i^T w + b\right) \geq 1 - \xi_i \tag{5-21}$$

El problema ahora es minimizar el sumatorio $\sum_{i=1}^{N}\xi_i$ a la vez que se minimiza el módulo de **w**. El hiperplano óptimo es el que minimiza el función:

$$\|w\| + C\sum_{i=1}^{N}\xi_i \tag{5-22}$$

donde C es un término de penalización de los errores ξ_i. Esto es equivalente a minimizar la función [Vapnik, 1995]

$$W(\alpha) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j d_i d_j x_i^T x_j \tag{5-23}$$

con la restricción

$$0 < \alpha_i < C \tag{5-24}$$

La máquina de clasificación anterior es lineal. La generalización a máquinas de clasificación con frontera no lineal se hace viendo la función no lineal como una combinación lineal de vectores imagen de los datos en un espacio intermedio. Este espacio es la salida de un conjunto de funciones no lineales cuya entrada son los datos.

Schölkopf et al. [Schölkopf, 1997] utilizan un mapeo de los datos mediante una transformación no lineal con RBF:

$$z_j = \left\{ e^{\frac{\|x_i - x_j\|}{2\sigma}} \right\}, \quad 1 \leq i \leq N \tag{5-25}$$

Es decir, se pasan todos los datos a través de un conjunto de RBF gaussianas construidas con todos los datos como centroides. En este espacio se puede definir un hiperplano de separación como

$$\sum_{i=1}^{N} w_i e^{\frac{\|x - x_i\|}{2\sigma}} + b = 0 \tag{5-26}$$

La optimización se puede aplicar a este espacio, obteniendo **w** y $\mathbf{x}_i$, que son el vector de pesos de la capa de salida y los centroides de una red RBF. Esta

red RBF proporciona la separación óptima del conjunto de datos para un valor determinado de σ.

Esta técnica ha demostrado buenos resultados en diferentes problemas de clasificación. Sus principales inconvenientes son, a nuestro entender, tres:

- el coste computacional necesario para construir la máquina es elevado, ya que hay que resolver un problema de minización cuadrática con restricciones;
- la inherente no adaptabilidad del proceso. Si los datos del problema de clasificación tienen una estadística no estacionaria (como en igualación de canal) la máquina SVM debe ser recalculada cada vez que el canal cambia su respuesta al impulso. Como el sistema ha de recoger datos para reconstruir la máquina, la transmisión de datos ha de ser interrrumpida para pasar a transmitir un protocolo de entrenamiento con suficientes datos para construir una máquina que presente baja tasa de error. Por supuesto, la longitud mínima del protocolo depende de las características de los datos, del canal y del ruido. La construcción de un igualador adaptativo basado en SVM es, en el momento de escribir esta tesis, una cuestión abierta.
- Una característica de las máquinas SVM es que cuando los datos presentan un alto solapamiento, el método selecciona un alto número de ellos, incrementando el coste computacional de la clasificación. Burgues [Burgues, 1997], [Schölkopf, 1997] demostró que la máquina SVM puede aproximarse por otra construida con un subconjunto menor de sus vectores de soporte (vectores de soporte generalizados). Sin embargo, esto incrementa el coste computacional de la construcción de la máquina.

El método propuesto en este trabajo está inspirado en la máquina SVM. Sin embargo, en lugar de utilizarse el principio de minimización del riesgo estructural (principio en el que se basan las máquinas SVM descritas arriba), se utilizan criterios de selección de muestras críticas. Se utilizarán como instrumentos los métodos de selección de centroides y de estimación de desviaciones típicas descritos en este mismo capítulo.

Con ello se ha conseguido construir una máquina tipo SVM que tiene buenas propiedades de adaptación, seguimiento de canal no estacionario y de tasa de error [Martínez/1, 1999]. Pero además se ha conseguido construir una máquina [Lyhyaoui, 1999] que, con muchos menos vectores de soporte, consigue mejor generalización que la máquina SVM original en los problemas de clasificación en los que se ha utilizado.

5.4.2. Concepto de muestra crítica

En nuestra aportación el enfoque es el mismo en cuanto que se utiliza una serie de vectores de soporte para construir una máquina de clasificación que luego es entrenada con las otras muestras del conjunto de entrenamiento. La

diferencia es la siguiente: los vectores de soporte se buscan de forma continua y se van sustituyendo por otros nuevos a medida que la estadística de los datos varía. Esto permite que el método sea adaptativo. El método SVM utiliza optimización no lineal para encontrar, en forma automática, los vectores de soporte: no necesita de una definición de vector significativo o vector crítico. En cambio, la variante que nosotros proponemos necesita de una definición de muestra crítica.

Una manera de seleccionar muestras críticas en forma adaptativa es examinar cada muestra en el momento de su recepción y, siguiendo un criterio de importancia basado en una frontera de referencia, se selecciona o no esta muestra.

La mayoría de los procedimientos de selección de muestras seleccionan como muestras criticas aquéllas que producen los mayores errores [Cachin, 1994] [Zhang, 1994]. No obstante, estos métodos no están bien condicionados en problemas con muestras muy ruidosas. Además, esos criterios no proporcionan una manera sencilla de determinar el número de muestras que debe seleccionarse, lo que es un aspecto muy importante a tener en cuenta a la hora de construir un clasificador tipo SVM. Por otra parte, otras muestras que no tienen un error demasiado alto pueden ser importantes para construir el clasificador. Nuestra filosofía sigue la de [Munro, 1992], pero utilizando la clusterización como un paso de simplificación del algoritmo.

Para entender las razones por las cuales escogemos el método propuesto, se exponen las siguientes definiciones.

5.4.2.a Problemática de la selección de muestras basada en frontera de referencia

La definición más sencilla de muestra crítica es la siguiente: una muestra es crítica si, estando bien clasificada, está muy cerca de la frontera de referencia.

Si se escogen las muestras siguiendo esta definición de muestra crítica puede pasar que las muestras seleccionadas se acumulen en el punto de mínima distancia entre clases diferentes, haciendo que la frontera clasifique muy mal fuera de ese entorno. Esto, en particular, pasará cuando las muestras se estén distribuidas en agrupaciones ("clusters"). En ese caso, hay que tomar muestras que pertenezcan a diferentes agrupaciones, que no estarán necesariamente cerca de la frontera

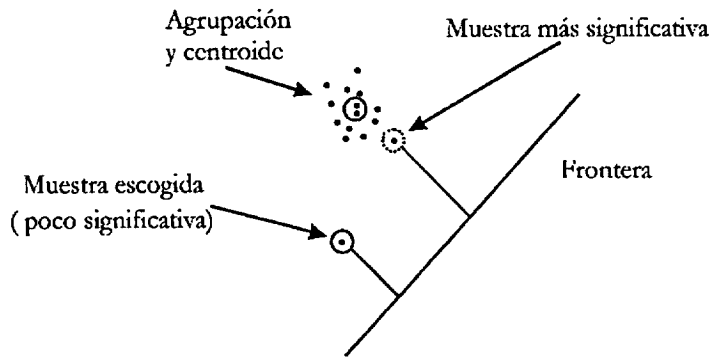


Figura 5-4 Si se utiliza sólo la distancia a la frontera como criterio para seleccionar muestras críticas dentro de un cluster, algunas de las muestras escogidas pueden tener una probabilidad de ocurrencia baja, o ser atípicas aún estando cerca de la frontera, no siendo, por tanto, realmente significativas.

La definición de muestra crítica pasa a ser la siguiente: una muestra es crítica si, estando bien clasificada, es la más cercana a la frontera de todas las que pertenecen a cada una de las agrupaciones de datos. Para encontrar las agrupaciones, debemos ayudarnos de una cuantificación vectorial que las identifique.

Por supuesto, no todas las agrupaciones serán importantes para la elaboración de la frontera de clasificación: hay que encontrar las que sean críticas mediante algún procedimiento. Cada agrupación de datos estará señalada mediante un centroide de la cuantificación vectorial. Entonces, aplicando el método de identificación centroides críticos, se encuentran las agrupaciones críticas.

Este criterio mejora el problema del anterior, pero tiene un fallo: si se presenta una muestra muy ruidosa comparada con las demás pertenecientes a un cluster determinado, de manera que está muy cerca de la frontera, tiene una posición muy sesgada, y resulta bien clasificada, será tomada como crítica por el anterior criterio (véase la Figura 5-4). Sin embargo, esta muestra será poco representativa y por tanto, se construirá con ella una frontera incorrecta. La definición de muestra crítica debe modificarse para salvar esta eventualidad.

5.4.2.b Selección de muestras basada en frontera de referencia y clusterización

Se encuentra la proyección de los centroides sobre la frontera de referencia. Si la cuantificación vectorial es representativa de la estadística de los datos, entonces estas proyecciones son muy importantes:

- dado un conjunto de muestras pertenencientes a un centroide y todas ellas a la misma distancia de la frontera, la más típica será la más cercana a la proyección del centroide.
- Por otro lado, dado un conjunto de muestras pertenecientes a un centroide

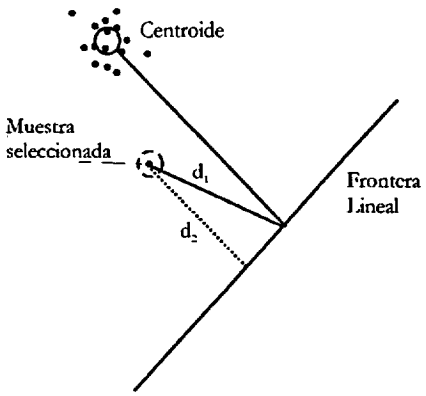


Figura 5-5. Criterio de selección de muestras: ponderación de las distancias d_1, desde la muestra a la proyección del centroide sobre la frontera, y d_2, desde la muestra hasta la frontera.

todas ellas a la misma distancia de su proyección, la más crítica será la más cercana a la frontera.

La selección de muestras, por lo tanto, se lleva a cabo teniendo en cuenta una combinación lineal de ambas medidas (Figura 5-5).

5.4.2.c Función indicadora

La función indicadora es una ponderación entre tipicidad de la muestra y proximidad a la frontera. La función elegida como indicadora es

$$I_1 = \|\mathbf{z} - \mathbf{z}_o\|_1 + \|\mathbf{w}^t\mathbf{z} + b\|_1 \qquad (5\text{-}27)$$

donde $\mathbf{z}$ es la imagen de los datos en un espacio de salida que puede ser el de los datos o una transformación no lineal del tipo

$$\mathbf{z} = K(\mathbf{x}, \mathbf{c}_j) \qquad (5\text{-}28)$$

La transformación elegida aquí es la función de base radial Gaussiana $K(\mathbf{x}, \mathbf{c}_j) = \exp\left(\|\mathbf{x} - \mathbf{c}_j\| / \left(2\sigma^2\right)\right)$, siendo $\mathbf{c}_j$ los centroides de una cuantificación vectorial y $\mathbf{z}_0$ la proyección de la imagen en el espacio de salida del centroide más cercano a la muestra $\mathbf{z}$ sobre la frontera ($\mathbf{w}$, b).

Para calcular la proyección del centroide más cercano a la muestra se toma la recta normal al plano y que pasa por el centroide $\mathbf{z}_j$:

$$\mathbf{z} = \lambda\mathbf{w} + \mathbf{z}_j \qquad (5\text{-}29)$$

Combinando esta ecuación con la del plano $\mathbf{w}^T\mathbf{z}+b=0$, el punto $\mathbf{z}_0$ de la recta contenido en el plano es aquél para el cual

$$\lambda = -\frac{\mathbf{w}^T \mathbf{z}_j + b}{\|\mathbf{w}\|^2} \tag{5-30}$$

Sustituyendo λ en la ecuación de la recta (5-29) se obtiene

$$\mathbf{z}_0 = \mathbf{z}_j - \frac{\mathbf{w}^T \mathbf{z}_j + b}{\|\mathbf{w}\|^2}\mathbf{w} \tag{5-31}$$

La función indicadora (5-27) presenta vértices, y esto puede acarrear inconvenientes cuando el número de dimensiones es elevado. Si se expresa la función indicadora (5-27) como

$$I_1 = \sqrt{\|\mathbf{z} - \mathbf{z}_o\|^2} + \sqrt{\|\mathbf{w}^t \mathbf{z} + b\|^2} \tag{5-32}$$

el gradiente de I respecto de x es (recuérdese que z es la imagen de x):

$$\begin{aligned} \nabla_x I_1 &= \left(\frac{\mathbf{z} - \mathbf{z}_0}{\|\mathbf{z} - \mathbf{z}_0\|} + \frac{\mathbf{w}^T \mathbf{z} + b}{\|\mathbf{w}^T \mathbf{z} + b\|}\mathbf{w} \right) \frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \\ &= \left(\frac{\mathbf{z} - \mathbf{z}_0}{\|\mathbf{z} - \mathbf{z}_0\|} + d\mathbf{w} \right) \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \end{aligned} \tag{5-33}$$

siendo d la clase de x. Este gradiente presenta una discontinuidad de salto sobre la frontera, donde d cambia de signo[22]. La indeterminación del tipo 0/0 del primer sumando de (5-33) tiene límite unidad.

Otras funciones, que varían la ponderación de tipicidad y proximidad a la frontera son

$$I_1(\beta) = \|\mathbf{z} - \mathbf{z}_o\| + \beta\|\mathbf{w}^t \mathbf{z} + b\| \tag{5-34}$$

o bien, ya que (5-27) y (5-34) son funciones que tienen vértices

$$I_2(\beta) = \|\mathbf{z} - \mathbf{z}_o\|^2 + \beta\|\mathbf{w}^t \mathbf{z} + b\|^2 \tag{5-35}$$

---

[22] Además, la solución de los mínimos de esta función en el espacio de entrada no es convexa: para cada mínimo en el espacio de z (mínimo que determina el carácter crítico de la muestra) puede haber más de un mínimo en el espacio de entrada de x. Estos mínimos pueden ser utilizados como criterio para seleccionar más de una muestra por centroide. En simulaciones previas en las que se han encontrado numéricamente estos mínimos, se ha visto que coinciden con las zonas en en las que existen muestras críticas. Mediante el método de selección expuesto, se encuentra el mínimo absoluto, despreciando los otros. El estudio de esta posible propiedad está fuera de los límites marcados en esta tesis.

Si se comparan (5-34) y (5-35) para $\beta = 1$ y para z_1 y z_2 tales que $I_1^2(z_1) = I_2(z_2)$ se obtiene

$$\begin{aligned} &\|z_2 - z_o\|^2 + \|w^t z_2 + b\|^2 = \\ &= \|z_1 - z_o\|^2 + \|w^t z_1 + b\|^2 + 2\|z_1 - z_o\|^2 \|w^t z_1 + b\|^2 \end{aligned} \tag{5-36}$$

Sobre la frontera $w^t z + b = 0$, por lo que $z_1 = z_2$. Fuera de ella, el tercer sumando de la derecha implica que z_1 está más cerca de la frontera que z_2: lo que significa que las ponderaciones de las dos ecuaciones son diferentes.

La función (5-27) tiene, no obstante, buenas prestaciones en selección de muestras. Si se opta por (5-35), para que la ponderación entre tipicidad y cercanía a la frontera sean lo más parecidas posibles, se escoge aquel valor de β que hace que la distancia entre las dos funciones sea la mínima. Para ello se fuerza a que las dos funciones se toquen en el punto z' que corta con la recta (5-29) normal a la frontera que pasa por el centroide y en la frontera:

$$\begin{aligned} &I_2(\beta) = I_1 \\ &\|z' - z_o\|^2 + \beta\|w^t z' + b\|^2 = \\ &= \|z' - z_o\|^2 + \|w^t z' + b\|^2 + 2\|z' - z_o\|^2 \|w^t z' + b\|^2 \end{aligned} \tag{5-37}$$

y de aquí

$$\beta = 1 + 2\frac{\|z' - z_o\|}{\|w^t z' + b\|} \tag{5-38}$$

Pero como z_0 es proyección de z'

$$z_0 - z = \frac{w^t z' + b}{\|w\|^2} w \tag{5-39}$$

Sustituyendo (5-39) en (5-38) y aplicando el resultado (5-35) en se obtiene la expresión siguiente para la función indicadora I_2:

$$I_2(\beta) = \|z - z_o\|^2 + \left(1 - \frac{2}{\|w\|}\right)\|w^t z + b\|^2 \tag{5-40}$$

Esta función produce resultados similares a las de I_1 evitando la aparición de vértices.

5.4.2.d Método de selección de muestras (SM)

La selección de muestras se hace de la siguiente manera:

- Se cuantifican vectorialmente los datos sobre el espacio de entrada.
- Se seleccionan de centroide críticos
- Para cada muestra nueva, se busca el centroide más cercano y se calcula el valor de I para esa muestra. Si el centroide no tiene una muestra asignada, se le asigna ésta. Si el centroide ya tenía una muestra asignada, se calcula su I. Se conserva la muestra que tenga I más baja.

5.4.3. Igualadores tipo SVM basado en Selección de Muestras

5.4.3.a Construcción del igualador

El igualador consta de un esquema en escalera, que agrupa los datos de maera que se puedan clasificar mediante una cadena de decisiones binarias. En las zonas del espacio que presenten mayor riesgo bayesiano, se inserta un clasificador basado en SM.

Al principio, se entrena un conjunto de centroides críticos según se detalla en 5.3.2.a, 5.3.2.b y 5.3.2.c. Simultáneamente a la clusterización se aplica el algoritmo de selección de muestras a cada uno de los centroides:

- se entrena una frontera de referencia: puede ser construida con los centroides o bien, en algunos casos (canal de fase mínima), la frontera lineal del igualador en escalera produce bajas tasas de fallo junto con alta velocidad de convergencia y, por lo tanto, de selección de muestras;
- cuando se adquiere una nueva muestra, se determina a qué centroide pertenece;
- si el centroide no tiene muestra crítica asignada, se le asigna la presente. En caso contrario, se mide I para las dos muestras y se retiene como crítica la que presente un valor menor, rechazando la otra.
- Con el conjunto de muestras se construye una red RBF, cuya capa de salida se entrena mediante un algoritmo de gradiente

En el instante de eliminación de centroides críticos, las muestras asociadas no se eliminan inmediatamente, sino que se conservan para verificar si alguna de ellas puede ser muestra crítica de alguno de los centroides restantes, lo que tiende a acelerar el proceso.

5.4.3.b Selección de varias muestras por centroide

En algunas situaciones, la frontera de clasificación determinada por la selección de una muestra por centroide es insuficiente. Es el caso en que un centroide de una clase está rodeado por varios centroides de clase contraria. El método de selección de una sola muestra escogerá la que esté más cerca del

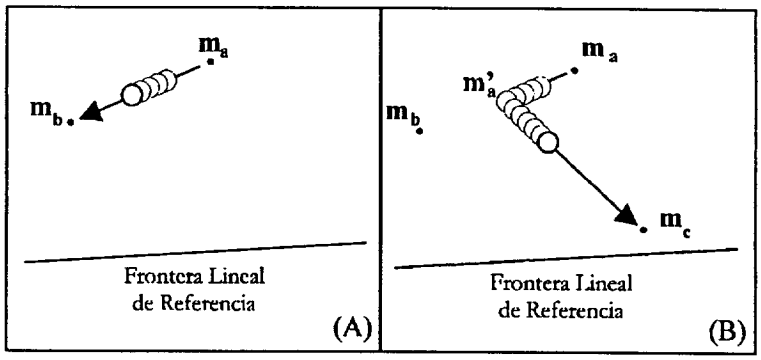


Figura 5-7 (a) Transición suave desde una muestra crítica antigua a otra nueva. (b) Si durante la transición aparece una muestra nueva, se toma como muestra antigua m'_a la combinación lineal de las muestras m_a y m_b en ese instante y se reinicia el proceso con λ=0.

mínimo global de la función indicadora. Pero en esta situación habrá más mínimos locales[23], que corresponden a muestras críticas cercanas a los centroides de clase contraria. Ante la dificultad (no resuelta) de encontrar analíticamente estos mínimos, se opta por seleccionar varias muestras por centroide. Para evitar que todas las muestras estén unas al lado de las otras en torno al mínimo global (lo que pasará siempre a la larga) se introduce una dispersión en la heurística de selección: una vez se ha seleccionado una muestra para un centroide, se puede seleccionar otra nueva si la distancia a su centroide es menor a la distancia entre la nueva muestra y la anterior. Si se cumple esto, se selecciona la nueva muestra. Si no, se selecciona la que tenga menor I y se rechaza la otra. El mismo principio puede ser aplicado a un número mayor de muestras.

5.4.3.c Combinación lineal de muestras

El igualador substituye muestras antiguas por otras nuevas que sean más críticas durante todo el proceso. Como la RBF que se sitúa sobre las muestras nuevas cambiará su posición manteniendo el peso antiguo hasta que el entrenamiento por gradiente lo actualice, la adaptación de todo el sistema se puede volver ruidosa. Normalmente esto pasará sólo al principio o cuando haya un cambio en el canal. Después de un cierto periodo las muestras nuevas estarán muy cerca de las antiguas y el sistema no se perturbará significativamente. Si se observa que la transición es ruidosa, puede optarse por una transición suave de la muestra antigua m_a a la nueva m_b según

$$m = (1 - \lambda[n])m_a[n] + \lambda[n]m_b[n] \qquad (5\text{-}41)$$

con $\lambda[n]$ creciendo proporcionalmente a n desde 0 hasta 1 (Figura 5-7.a), en tanto que se entrena la red simultáneamente. Si durante la transición de una

[23] Véase nota 22

muestra a la otra aparece una muestra más crítica (Figura 5-6.b), debe reiniciarse el proceso tomando como muestra antigua la combinación lineal de ambas muestras en ese instante.

5.4.3.d Seguimiento de canales no estacionarios

La red RBF basada en selección de muestras tiene una ventaja como igualador de canal respecto de la red RBF basada en centroides: aquélla tiene un mejor seguimiento de canales no estacionarios.

Los centroides se actualizan continuamente, esto es, se mueven de acuerdo con la posición de las muestras que se adquieren. Esto les da su carácter adaptativo en canales no estacionarios. Pero, por otro lado, esta actualización produce un ruido de desajuste en la frontera: cada vez que se actualiza un centroide la frontera se mueve. Cuanto mayor sea el paso de adaptación de los centroides, mayor la capacidad de seguimiento y mayor también el ruido de desajuste: en la red de centroides hay que establecer una solución de compromiso entre ruido de desajuste y rapidez.

Si se utiliza una red RBF basada en selección de muestras, el ruido de desajuste sólo se produce cuando la muestra que llega es más crítica que una de las anteriormente seleccionadas, lo que es poco probable una vez se ha llegado al estado estacionario. Hay que resaltar dos aspectos:

- los centroides utilizados en la selección de muestras, al no intervenir directamente en la construcción de la frontera de clasificación, pueden tener pasos de adaptación mucho mayores, con lo que se adaptarán más rápido a cambios de canal: aunque los centroides tengan un movimiento dentro de sus clusters, las muestras críticas no se moverán, excepto que llegue una muestra más critica, y siempre que el movimiento sea de pequeña amplitud comparada con la desviación típica de los datos dentro del cluster;
- si hay un cambio en la estadística de los datos, esto es, el canal varía, la frontera de referencia tenderá a seguir las variaciones del canal: la probabilidad de que lleguen muestras más críticas que las actuales aumenta bruscamente. Es decir, algunas de las muestras críticas antiguas quedarán mal clasificadas, con lo que se sustituirán inmediatamente; otras quedarán muy lejos de la frontera, y, por tanto, rápidamente llegarán muestras más críticas. Cuando pase esto, la frontera construida con las muestras selecionadas sufrirá un reajuste "grueso" con la llegada de relativamente pocas muestras. Luego, mediante el entrenamiento por gradiente, la frontera llegará al ajuste fino.

Estos dos efectos hacen que (en la práctica) la frontera del igualador basado en muestras se mueva tan rápidamente como la frontera de referencia, o más en algunos casos. Si la frontera de referencia es lineal, la convergencia será muy rápida (véase [Widrow, 1984] para una discusión de las prestaciones del LMS en canales no estacionarios). Si se utiliza la frontera RBF basada en los centroides

críticos, se puede emplea un paso de adaptación mayor que si se utilizara esta frontera para clasificar, ya que el ruido de desadaptación es poco importante, y así se consigue mejor seguimiento.

5.4.3.e Elección de la desviación típica de las funciones de base radial

En 5.3.2.c se ha expuesto una técnica para estimar los valores de σ_i en las RBF. Pero esta aproximación no es válida cuando las RBF están centradas en muestras críticas. La estimación de los valores de las σ_i se basa en que los centroides están situados aproximadamente en máximos de densidad de probabilidad. La estimación se hace basándose en una aproximación Gaussiana de la distribución de los datos en torno al centroide. Cuando se trata de muestras críticas, lo anterior no es válido, ya que estas muestras están más bien cerca de las colas de las distribuciones cuyo centro es el centroide. Haciendo esta suposición, es posible hacer una estimación de las σ_i para las RBF de las muestras a partir de aquellas de los centroides.

La idea principal es la misma: se sitúa la zona de mínima sensibilidad de la frontera a una cierta distancia de la muestra, por debajo de la cual se puede suponer que se encuentra la mayoría de los datos pertenecientes al centroide.

Utilizando el mismo razonamiento que en 5.3.2.c se determina que para muestra crítica, $\sigma = \delta d'$, donde d' es la semidistancia entre la muestra para la que se pretende estimar σ y su muestra crítica de clase contraria más cercana, y δ es el factor de proporcionalidad

$$\delta = \frac{\sqrt{2}}{2}\sqrt{\frac{k_x^2 + d'^2}{d'^2}}$$

donde $k_x = \sqrt{3}d$ siendo d la semidistancia entre centroides. Por lo tanto, y considerando $d >> d$:

$$\delta = \frac{\sqrt{2}}{2}\sqrt{\frac{3d^2 + d'^2}{d'^2}} \approx \sqrt{\frac{3}{2}}\frac{d}{d'} \qquad (5\text{-}42)$$

5.4.3.f Muestras erróneamente seleccionadas

Las clases pueden llegar a estar, algunas veces, altamente solapadas a causa del ruido. Por otra parte, más si se utiliza como referencia una frontera lineal, la clasificación de los datos (téngase en cuenta que el entrenamiento es dirigido por decisión) hará que se tomen algunas muestras como críticas cuando en realidad están del lado incorrecto del hiperplano de separación óptimo: esto es, la frontera de referencia clasifica correctamente una muestra que quedaría mal clasificada si

el hiperplano de clasificación generalizase mejor. Si se da esta situación. la frontera que se construya no generalizará correctamente. Se pueden aplicar dos soluciones a este defecto:

- introducir en la función indicadora un término adicional:

$$I_1(\beta) = \|z - z_o\|_1 + \beta\left(\|w^t z + b\|_1 - \xi\right) \qquad (5\text{-}43)$$

donde $\xi > 0$. Es fácil ver que el mínimo de esta función está a una distancia $\xi/\|w\|$ de la frontera de referencia, lo que disminuye la probabilidad de escoger una muestra falsamente crítica. En realidad, esta restricción se impone en el algoritmo SVM para conjuntos no separables linealmente, en la ecuación (5-21) [Vapnik, 1995]. Esta restricción disminuye el número de muestras clasificables, que es lo que se busca, pero como contrapartida disminuye la velocidad de convergencia. La otra posible solución es

- utilizar una combinación lineal entre el centroide y la muestra:

$$m_i' = \upsilon c_i + (1 - \upsilon) m_i \qquad (5\text{-}44)$$

con $0 < \upsilon < 0.8$. La cota superior es experimental. A m_i' la llamamos pseudomuestra, ya que en realidad es un punto en el espacio no ocupado por ningún dato.

5.4.4. Igualador SM para señales binarias

5.4.4.a Convergencia

Se toma un conjunto de 5000 símbolos binarios que pasan a través del canal h[n] = 1 + 0.5δ[n-1]. Los datos son de orden dos.

Se han hecho ensayos del comportamiento del algoritmo y se han comparado los resultados del método de Selección de Muestras con los resultados del algoritmo SVM (esta comparación se hace tomando las 5000 muestras y aplicando el algoritmo de minimización cuadrática (5-23) para funciones de base radial).

El número de clusters para este caso particular es de 8. El número inicial de centroides es 36 y la conmutación a decisión dirigida por datos se hace a partir de la muestra número 100.

Al haber poco solapamiento de clases, se utiliza la función indicadora

$$I_1(\beta) = \|z - z_o\|_1 + \|w^t z + b\|_1$$

En todas las simulaciones se utiliza sustitución directa de muestras, es decir, no se utiliza la combinación lineal de muestras de (5-41).

La clusterización seguida de la selección de centroides críticos siempre escoge 8 centroides, eliminando todos los demás. Ocuparán los posiciones de los 8 clusters del espacio de datos. Se seleccionan, por tanto, 8 muestras. En la Figura 5-7 puede verse una realización del algoritmo. En ésta, los centroides críticos son los marcados con las etiquetas 1, 2 y 3, mientras que los marcados con la etiqueta 4 no lo son. En estas simulaciones se utilizan sólo los centroides críticos, pero no se eliminan los no críticos, ya que su índice de utilización U_i cae por encima del umbral.

Los resultados para diferentes relaciones de señal a ruido son los siguientes:

- SNR=10 dB: En la Figura 5-8 puede verse la frontera SM, las muestras críticas (círculos unidos por una línea discontínua a la proyección de los centroides), y los vectores de soporte de SVM (aspas y cruces). El método SVM escoge un gran número de muestras, muchas de las cuales están mal clasificadas por la frontera. Se observa un aparejamiento de muestras de las dos clases (que podrían ser eliminadas en la construcción de la frontera por ser nula su contribución). El algoritmo SM escoge 8 muestras,aunque sólo 6 corresponden a centriodes críticos. Las muestras escogidas por el método SM no coinciden con muestras del método SVM. Las prestaciones en cuanto a error de clasificación son muy parecidas, siendo algo mejores las de este ultimo método, aunque no parece justificable su coste computacional a la vista de los resultados.

- SNR=13 dB: El resultado de una realización se ve en la Figura 5-9. El algoritmo SVM escoge algunas muestras en el centro del espacio de datos, que es donde éstos están más cerca de la frontera. Escoger más muestras de los centroides que estén más cerca de la frontera puede hacer que la frontera sea mejor.
- SNR=16 dB: Los resultados de ambos algoritmos (Figura 5-10) son idénticos en prestaciones. El algoritmo SVM selecciona 4 muestras que coinciden con las cuatro más cercanas a la frontera de las 6 del algoritmo SM.

En la Figura 5-11 se muestran las gráficas comparadas de las tasas de error del algoritmo SM con las de la frontera óptima de Bayes y la frontera LMS.

En la Figura 5-12 se muestran las gráficas de U_i para los ocho centroides finales (gráfica superior) y de los instantes de eliminación de centroides no utilizados o con U_i. muy por debajo de la media. Se ve cómo la mayoría de los centroides se eliminan antes de las 300 muestras. La eliminación de centroides se activa después de las 150 primeras muestras. En ese instante, y para esta realización, se eliminan simultáneamente 11 centroides no utilizados, quedando tan sólo 25.

Desde ese instante hasta la muestra 320 se eliminan 14 centroides más, restando 11. Unas 275 muestras más tarde, se ha eliminado otros 3 centroides, quedando 8, número que coincide con el número de clusters. La selección de centroides críticos hace que sólo 6 muestras de esos 8 centroides se utilicen.

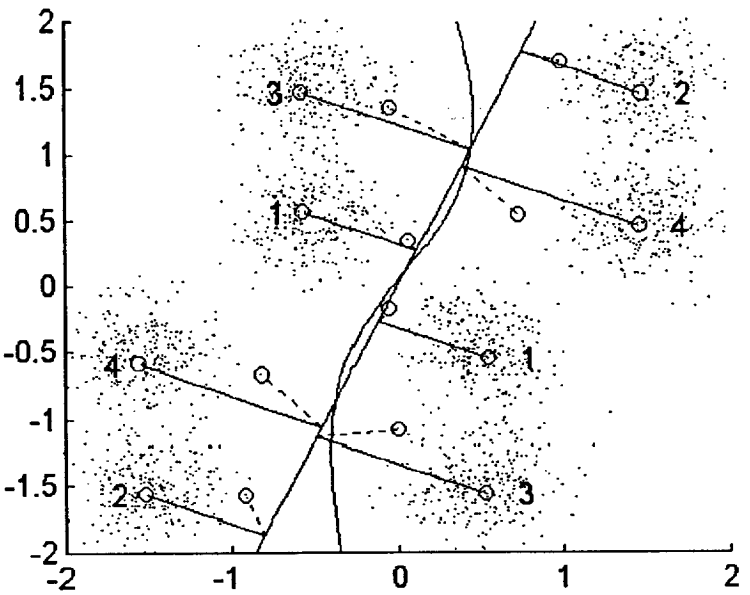


Figura 5-7. Realización del algoritmo SM. Se muestran los ocho centroides correspondientes a los ocho clusters y las muestras seleccionadas para cada uno de ellos. Los críticos son los marcados como 1, 2 y 3, que son los que finalmente se utilizan en la construcción de la frontera.

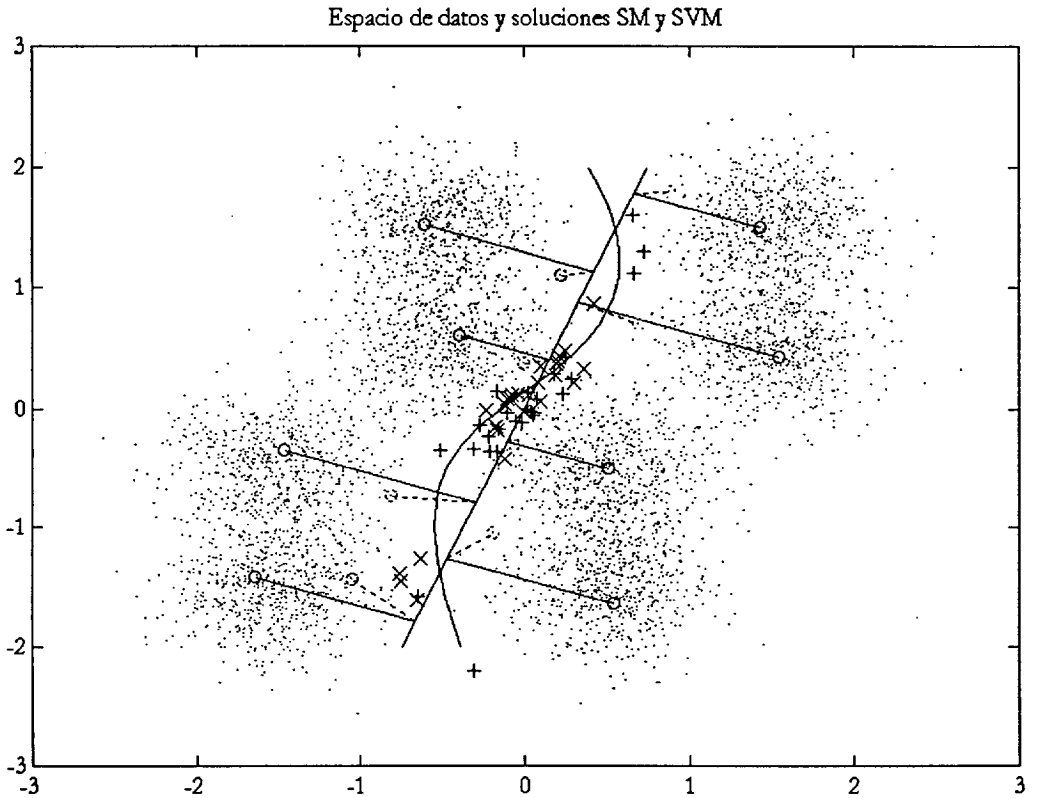


Figura 5-8. Comparación de los algoritmos SM y SVM en igualación de señales binarias para una SNR de 10 dB. El eje vertical representa las señales en el instante [n-1] y el horizontal en el instante [n]. El algoritmo SM escoge ocho muestras (círculos cercanos a la frontera), mientras que el SVM selecciona unos 40 ('×' para la clase de la derecha u '+' para la clase de la izquierda). Obsérvese que muchas de las muestras SVM están mal clasificadas. En esta realización el número de muestras mal clasificadas por el SM es 16, mientras que el SVM clasifica mal 15 muestras (en esta realización). La frontera corresponde a la solución proporcionada por el algoritmo SM. Las muestras escogidas por SM no son un subconjunto de las SVM. Los centroides críticos son los mismos que en la Figura 5-7.

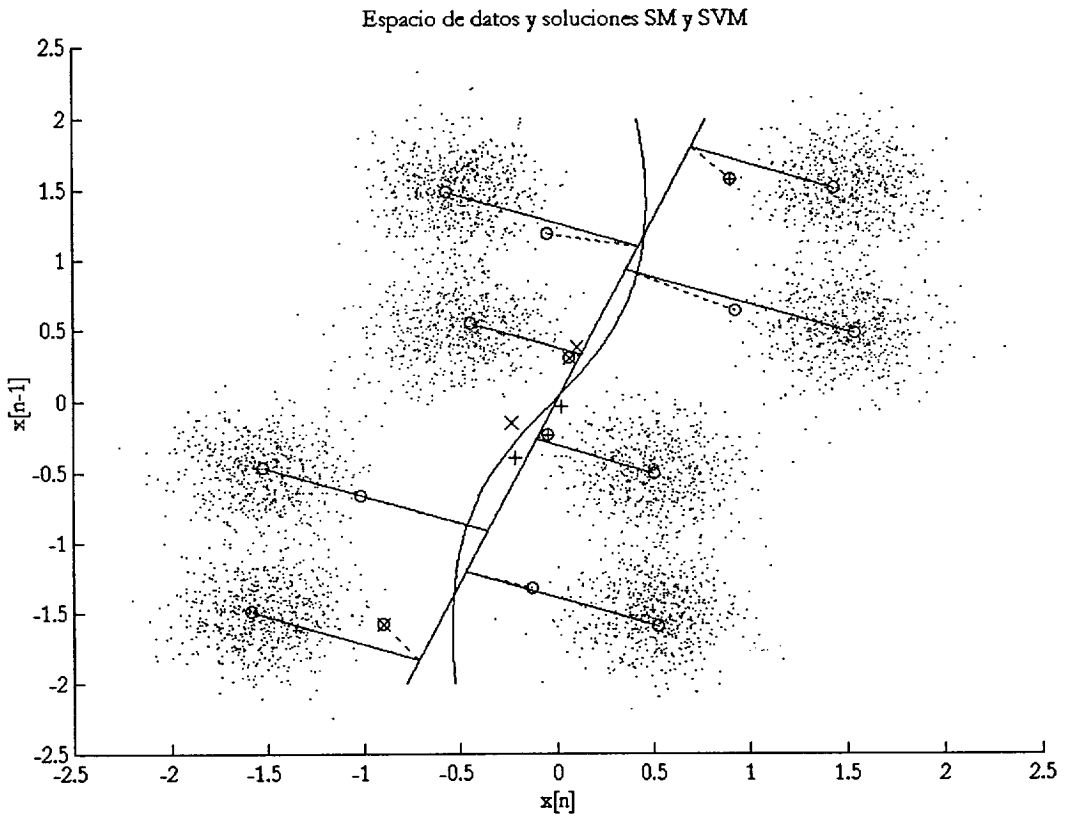


Figura 5-9. Misma comparación que en la Figura 5-8, pero para SNR=13 dB. Para tal relación de señal a ruido las muestras del algoritmo SM son un subconjunto de las muestras del algoritmo SVM.

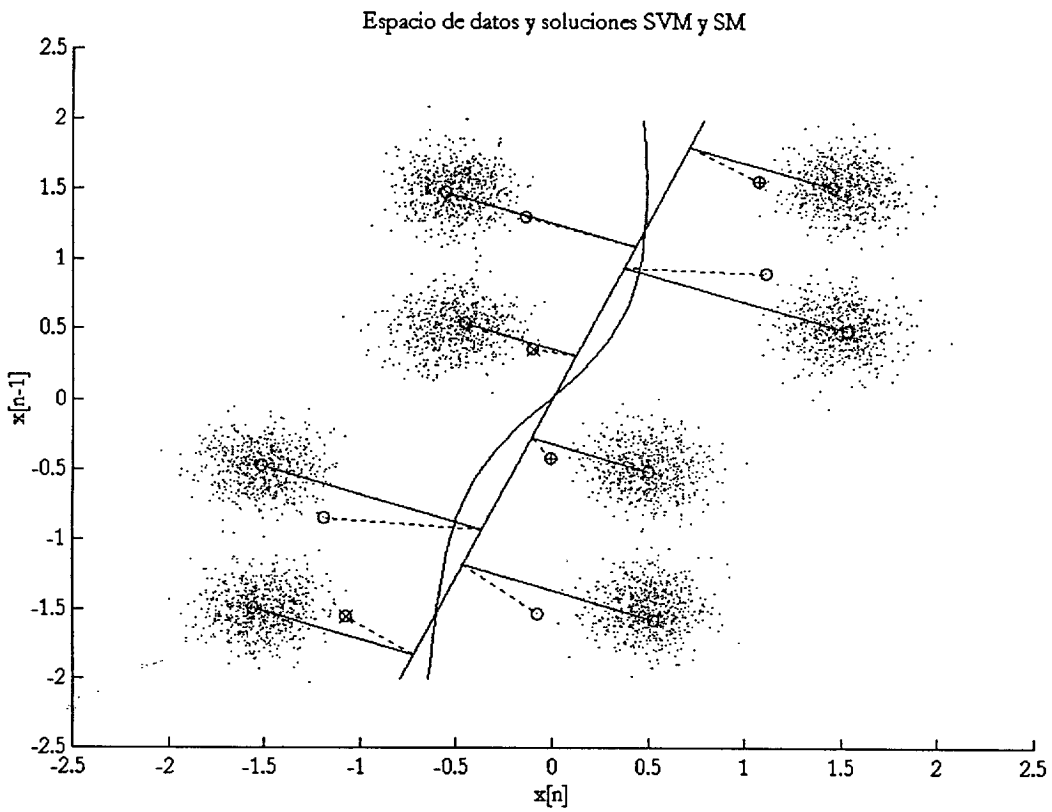


Figura 5-10. Misma comparación que las dos anteriores para SNR=16 dB. En este caso, los cuatro vectores seleccionados por el algoritmo SVM se corresponden con los cuatro más cercanos a la frontera de los ocho que selecciona el algoritmo SM: las muestras SVM son un subconjunto de las SM.

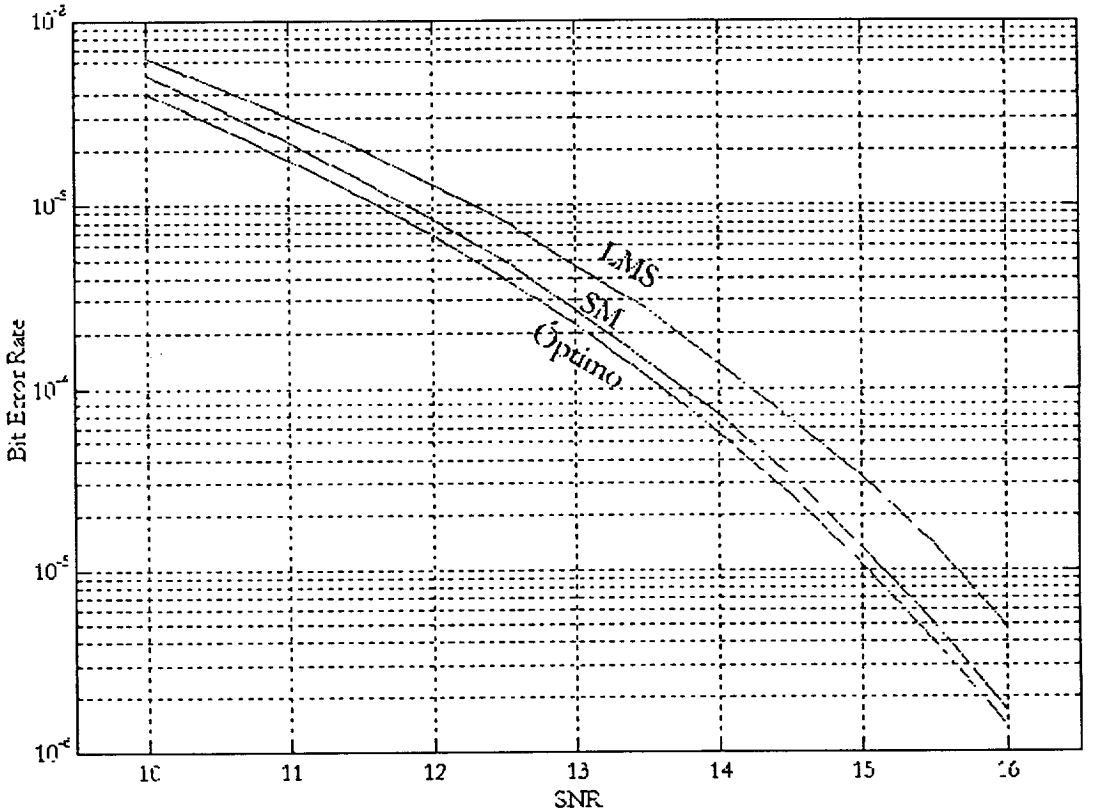


Figura 5-11. Comparación de tasas de fallo del algoritmo SM, el LMS y la frontera óptima. Para relaciones de señal a ruido pobres la frontera lineal y la óptima se parecen mucho por cuanto las gaussianas centradas en cada uno de los clusters de datos son muy anchas, lo que hace que la frontera óptima sea muy suave. Para altas relaciones de señal a ruido la frontera optima es muy alineal. La frontera SM se acerca razonablemente a la frontera óptima.

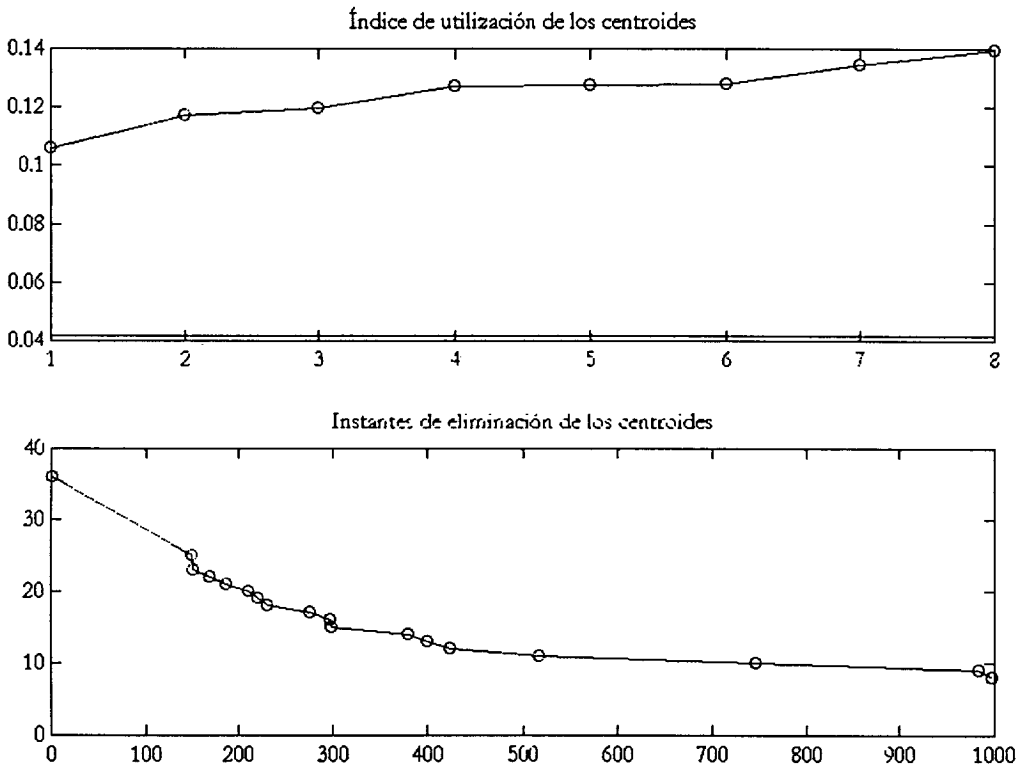


Figura 5-12. Indice de utilización de los 8 centroides que han quedado después de eliminar los demás (parte superior de la figura) en una realización con SNR=10 dB. La utilización de cada centroide después de 5000 muestras está entre el 10 y el 14%. En la parte inferior de la figura se observan los instantes de eliminación de los centroides. La velocidad de eliminación de centroides es mayor cuanto mayor es la SNR. Para SNR=16 db se eliminan todos los centroides sobrantes antes de las 1000 muestras.

5.4.4.b Seguimiento de un canal no estacionario

Se compara el algoritmo SM con el LMS. Se utiliza un canal con relación de señal a ruido de 12 dB. En el instante inicial el canal tiene la forma

$$h[n] = \delta[n] + 0.5\delta[n-1]$$

y el entrenamiento es guiado.

En el instante n=100 se conmuta a funcionamiento dirigido por decisión y, a los 250 datos se lleva a cabo la selección de centroides críticos. En el instante n=1500 se produce un cambio de la respuesta al impulso del canal a $h[n] = \delta[n] + 0.3\delta[n-1]$. Después de 500 muestras cuando n=2000, se produce otro cambio de canal a $h[n] = \delta[n] - 0.5\delta[n-1]$.

Se mide el error cuadrático medio (MSE) para el algoritmo LMS y se promedia para 100 realizaciones. Se ha verificado que en ninguna de las realizaciones se produzca un desenganche que pudiera falsear la medida.
El resultado se observa en la Figura 5-13. Se ve cómo la convergencia de los dos algoritmos es similar en velocidad, alcanzando el algoritmo SM niveles de error cuadrático medio más bajos que el LMS.

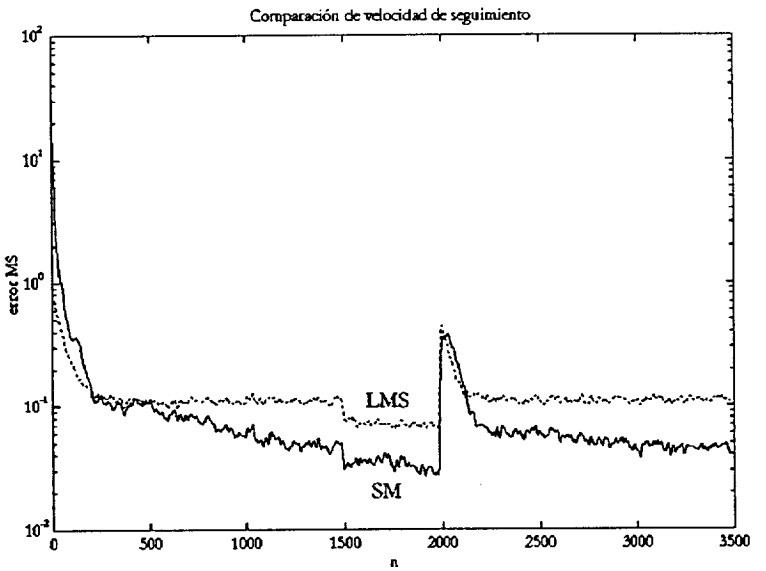


Figura 5-13. Seguimiento de canal no estacionario. La SNR es de 12 dB. En el instante inicial el canal tiene la forma h[n]= δ[n]+0.5·δ[n-1] y el entrenamiento es guiado. En el instante n=100 el sistema pasa a ser dirigido por datos. En el instante n=1500 el canal cambia a h[n]= δ[n]+0.3·δ[n-1] y en el instante n=2000 a h[n]= δ[n]-0.5·δ[n-1]. Se promedia sobre 100 realizaciones sin que en ninguna de ellas se haya producido desenganche.

5.4.5. Igualador SM para señales PAM

Se simula aquí el canal no lineal con memoria con las mismas características que en las anteriores simulaciones de igualadores PAM, pero, para observar mejor la capacidad de seguimiento del algoritmo, la respuesta al impulso del canal tiene una variación de mayor amplitud y es algo más suave:

- Respuesta impulsional de la parte lineal del canal:

$$h[n] = \delta[n] + h_1[n]\delta[n-1]$$

con

$$h_1[n] = \begin{cases} 0.2 & 1 < n < 3000 \\ 0.995h_1[n-1] - 0.001 & 3001 < n < 5000 \end{cases}$$

- zno linealidad $G(x) = \tanh(x/\xi)/\tanh(1/\xi)$ con $\xi = 7$.
- SNR= 30 dB

La señal emitida es 8-PAM con amplitudes {±1, ±3, ±5, ±7} ("back-off " de 0 dB).

El receptor es un esquema en escalera de orden 2 en el que el filtro para los símbolos {±5, ±7} es sustituido por un igualador RBF basado en muestras. El filtro no se elimina, sino que se utiliza como frontera de referencia.

Se inicializa distribuyendo uniformemente 15 centroides para cada una de las clases. Se estima que un conjunto de 250 datos es suficiente para entrenar los centroides (como es el caso en señales binarias). Para que, en media, la cantidad de datos de las clases {±5, ±7} sea de 250, el número total de datos debe ser 1000. Por lo tanto, la selección de centroides críticos lleva a cabo en el instante n=1000. El protocolo de entrenamiento guiado consta de un número total de 250 datos (los necesarios para que la frontera lineal de referencia converja).

El indicador utilizado aquí es

$$I_1(\beta) = \|z - z_o\|_1 + \beta\left(\|w^t z + b\|_1 - \xi\right)$$

con ξ=0.2.

Se representan gráficas del error cuadrático medio de la salida comparadas con las del algoritmo en escalera lineal y de red de RBF basada en centroides críticos. También se representa una gráfica comparativa de tasa de error en función de la SNR para un "back-off" de 0 dB.

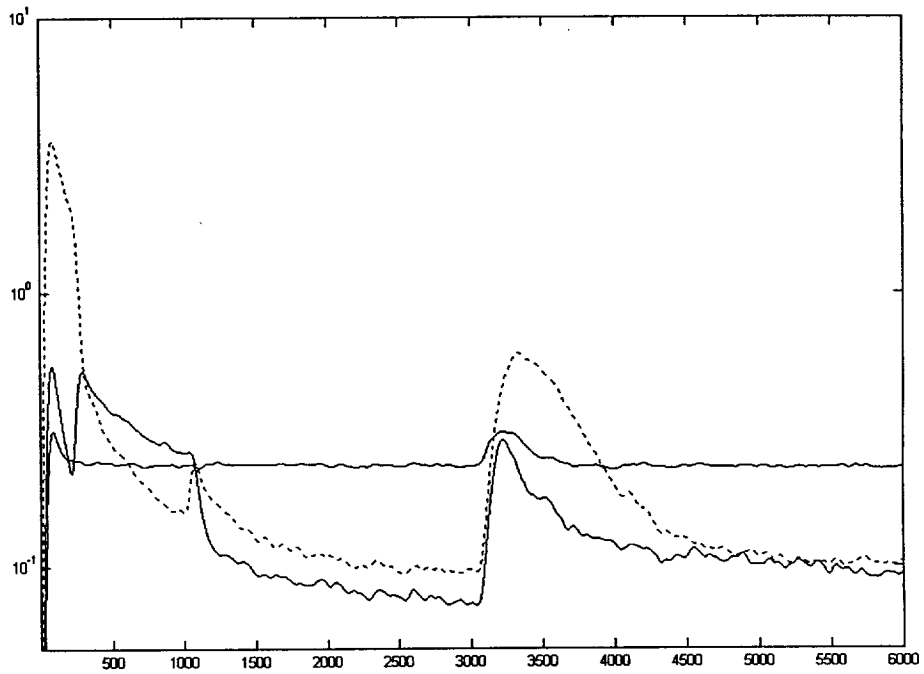


Figura 5-14. Comparación del error cuadrático en convergencia y para un canal no estacionario (a partir de la muestras número 3000) del filtro en escalera lineal (línea continua superior), la red RBF basada en centroides críticos (línea discontinua) y la red basada en muestras críticas (línea continua inferior). Los picos en torno a los instantes 250 y 1000 corresponden al inicio de la eliminación de centroides poco utilizados y al inicio de la selección de centroides críticos. Es importante tener en cuenta que las magnitudes absolutas del error cuadrático medio no tienen relevancia, por cuanto están medidas en espacios vectoriales distintos, sino las velocidades de convergencia. La simulción consta de 300 realizaciones y se ha suavizado el error con una ventana de Hamming de longitud 100.

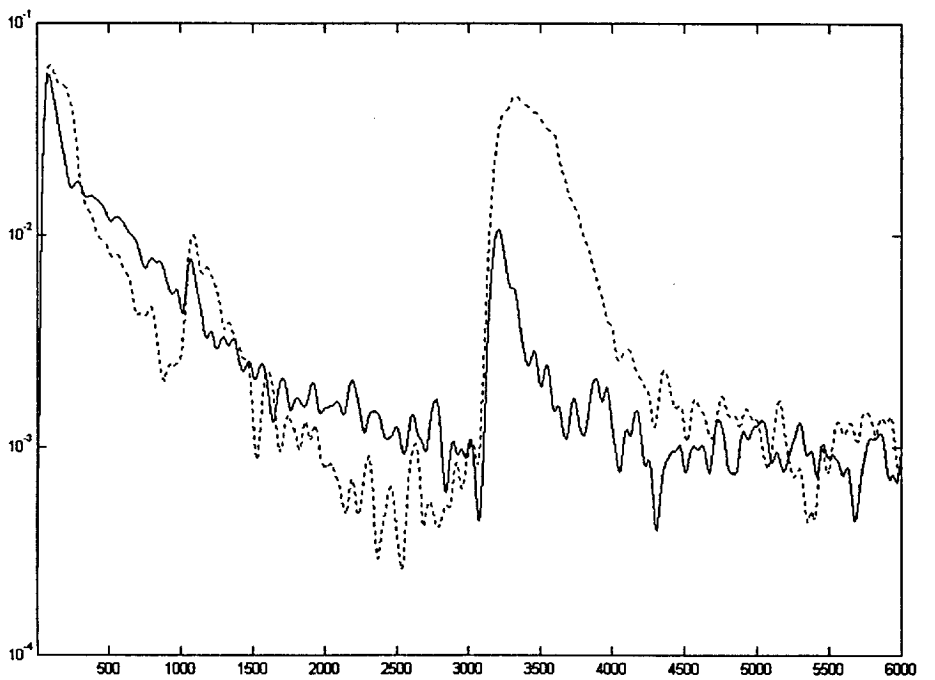


Figura 5-15. Tasas de error correspondientes a las simulaciones de la red de muestras (trazo continuo) y .la red de centroides (trazo discontinuo). Se ha suavizado con una ventana de Hamming de longitud 100.

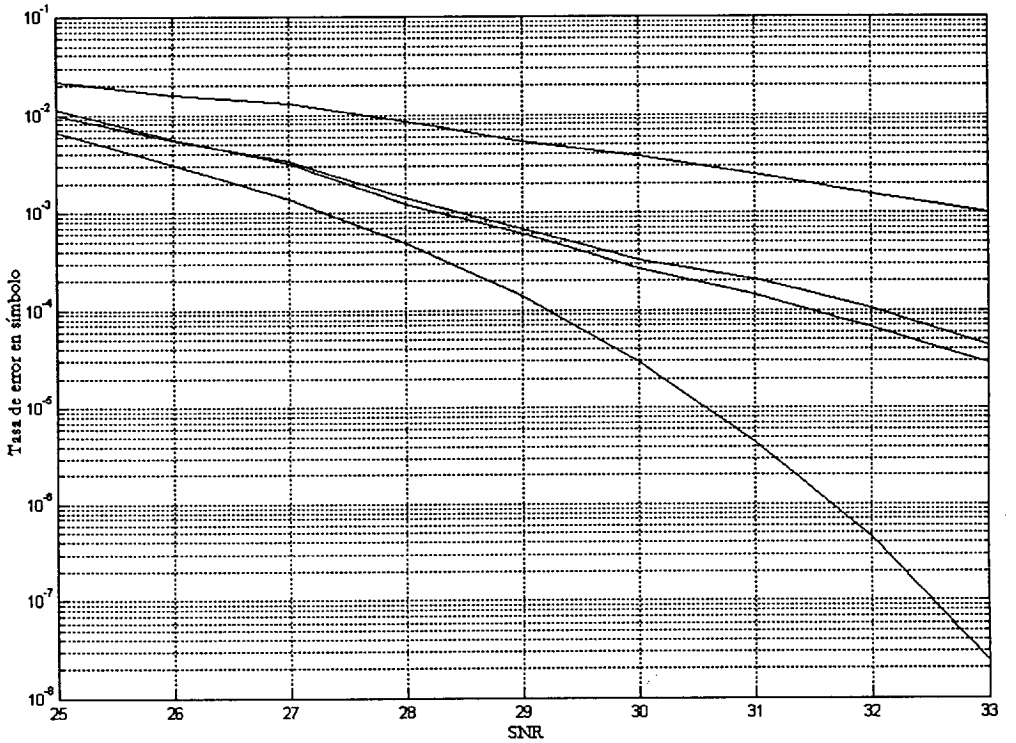


Figura 5-17. Comparación de las tasas de error del algoritmo en escalera lineal, el basado en polinomios de Volterra, Red de centroides críticos y red de Muestras.

5.5. DISCUSIÓN

En este capítulo se han presentado algunos esquemas no lineales basados en el esquema en escalera. Se pueden clasificar en dos tipos: aquellos que están construidos previamente a su puesta en marcha, como los polinomios de Volterra o los GCMAC, y aquellos que se construyen en función de los datos. Estos son la red RBF de centroides críticos y la red RBF de muestras críticas.

El algoritmo basado en muestras tiene una mayor velocidad de seguimiento que el algoritmo basado en centroides críticos, dado que el último conserva en la posición de los centroides la memoria de la estadística de los datos. Si la estadística cambia, los centroides se moverán lentamente hasta adaptarse a la nueva situación. En cambio, el algoritmo basado en muestras no tiene más memoria que la que conserva la frontera de referencia. Para los ejemplos vistos, la frontera puede ser lineal, con lo que su convergencia será rápida. El algoritmo de muestras críticas converge cuando se detectan nuevas muestras críticas, lo que sucede rápidamente cuando hay un cambio de canal. Se puede observar entonces un periodo de convergencia "gruesa" muy rápido, y un periodo de convergencia fina más lento, debido a que los centroides que se utilizan en la selección de muestras cambiarán su posición, y con ella el criterio de selección variará. Entonces, las muestras seleccionadas cambiarán ligeramente sus posiciones. Por otro lado, el conjunto de centroides puede permitirse tener un paso de adaptación alto, porque aunque tengan un ligero movimiento en el estado estacionario, ello no perturbará al sistema, ya que las muestras críticas no se mueven.

Como contrapartida, la red de centroides presenta, en estado estacionario, una tasa de error ligeramente más baja que la red de muestras, siempre que el paso de adaptación de los centroides sea suficientemente bajo. También puede observarse que la elección de sus parámetros es menos crítica: la variación de la anchura de las RBF en una red de centroides críticos produce cambios menos importantes que una variación igual en la red de muestras críticas.

El polinomio de Volterra produce peores prestaciones en todos los aspectos, aunque hay que remarcar que el algoritmo empleado es el más sencillo posible (lo que ya necesita un número de operaciones comparable al de los otros sistemas).

El algoritmo de muestras críticas presenta iguales prestaciones que las del algoritmo SVM en señales binarias, aunque esta comparación es puramente artificial: el algoritmo SVM no es adaptativo y las simulaciones se han hecho "off-line"

6. CONCLUSIONES Y LÍNEAS FUTURAS DE TRABAJO

6.1. CONCLUSIONES

Los esquemas de igualación no lineal en canales dispersivos producen fronteras de clasificación que disminuyen las tasas de error significativamente. Esto se traduce directamente la posibilidad de aumentar la velocidad de transmisión de datos. La ventaja de los sistemas no lineales frente a los lineales aumenta aún más cuando el canal presenta no linealidades.

Por otro lado, la igualación basada en esquemas no lineales tiene el inconveniente de que la complejidad de cálculo es, en general, elevada. Además, estos esquemas necesitan de un gran número de datos para poder llegar a la convergencia. Véase, por ejemplo, el perceptrón multicapa, que tiene el inconveniente añadido del riesgo de caer en mínimos locales. Los esquemas de Volterra tienen menos riesgo de caer en mínimos locales, pero son difíciles de ajustar para que converjan.

En algunas situaciones, no es realmente necesario utilizar esquemas lineales. Utilizados convenientemente, los filtros lineales producen resultados intermedios que constituyen una solución de compromiso entre tasa de error, velocidad de convergencia y complejidad.

Uno de estos esquemas es el algoritmo en escalera. Este algoritmo permite

clasificar los datos mediante una cadena de decisiones bit a bit. Ya que las decisiones son bit a bit, pueden introducirse funciones de coste con objetivos binarios. Uno de ellos es la medida de la entropía relativa o función de coste de Kullback-Leibler.

Las aportaciones de esta Tesis Doctoral se enumeran a continuación:

- Análisis del comportamiento del esquema en escalera para diferentes canales y funciones de coste.

  Se ha visto que la entropía proporciona una alta insensibilidad al paso de adaptación, lo que permite entrenar el algoritmo con pasos de adaptación elevados, que producen velocidades de convergencia elevadas. El inconveniente de esta función de coste es que es más inestable que el coste cuadrático cuando el canal es no estacionario.

- Establecimiento de un método para favorecer la convergencia del algoritmo con coste entrópico.

  La función de coste de entrópico se basa en la aproximación de que las salidas de los clasificadores son una medida de la probabilidad a posteriori, lo que no es una buena aproximación si los filtros no están inicializados adecuadamente.

  El filtro central del igualador en escalera sí lo está. Además, se comprueba que los valores finales de los pesos de los filtros del igualador serán muy parecidos incluso en condiciones de alta dispersión y distorsión no lineal. Eso se aprovecha para mejorar el entrenamiento, lo que permite la convergencia del algoritmo con función de coste de Kullback-Leibler.

- Inserción de igualadores no lineales binarios en el esquema en escalera.

  En esta tesis se estudian algunos tipos de igualadores no lineales que actúan sólo en determinadas zonas del espacio gracias a la estructura del algoritmo en escalera

- Desarrollo de versiones de la Máquina de Vectores de Soporte, basadas en un criterio de Selección de Muestras y aplicación a igualación no lineal.

  Este criterio permite que, a partir de las muestras seleccionadas, la máquina de clasificación se autoconstruya, en lugar de diseñarla a priori, y se basa en un criterio de selección de centroides críticos desarrollado recientemente.

  Se han hecho simulaciones para diferentes canales no lineales y

constelaciones reales

Se ha visto que el sistema es capaz de seguir variaciones rápidas del canal y tiene bajas tasas de error en constelaciones reales con distorsiones no lineales elevadas.

- Métodos de estimación de parámetros de anchura para funciones de base radial.

  El parámetro de anchura de las funciones de base radial que se utiliza en las máquinas de selección de muestras se estima mediante un criterio de sensibilidad de la frontera a variaciones de los pesos de la capa de salida, en lugar de recurrir al entrenamiento por gradiente, que tiene una baja velocidad de convergencia.

6.2. TRABAJOS FUTUROS

- Combinación de funciones de coste

  Una de las ventajas del algoritmo en escalera frente a otros esquemas no lineales es la posibilidad de utilizar funciones de coste binarias, como la medida de la entropía, que tiene buenas prestaciones en velocidad de convergencia, ya que su tasa de error es muy poco sensible al paso de adaptación: se pueden escoger pasos elevados para acelerar la convergencia. Pero si se hace esto, el sistema se vuelve inestable frente a canales no estacionarios. Por el contrario, el coste cuadrático es capaz de seguir variaciones bruscas de canal. Es posible que una combinación de los dos costes adecuadamente gestionada produzca bajas tasas de error, estabilidad y velocidad de convergencia.

- Detección fina de zonas del espacio con alto riesgo de Bayes

  Los esquemas propuestos en esta tesis fijan a priori las zonas del espacio que tienen mayor riesgo de Bayes para utilizar igualadores no lineales en ellas. Pero estas suposiciones no tienen porqué ser ciertas, ya que dependen del modelo de no linealidad del canal. Así por ejemplo, con la no linealidad utilizada y para las simulaciones de canales no lineales con memoria, la única zona que presenta riesgo de Bayes alto está situada en uno solo de los cuadrantes del espacio, mientras que el esquema no lineal utilizado se extiende

mucho más allá de aquél. Si el sistema es capaz de localizar las zonas del espacio que necesitan de igualación no lineal, la reducción de coste computacional puede ser significativo.

- Generalización del esquema basado en selección de muestras a constelaciones complejas prácticas

  La generalización del esquema en escalera a constelaciones complejas es sencillo, ya que su estructura se basa en filtros lineales. Pero no se puede decir lo mismo con los esquemas propuestos de selección de muestras: es necesario establecer las pautas a seguir en el diseño de esquemas de igualación en constelaciones complejas.

- Igualadores basados en selección de muestras que utilicen otros Kernels

  El esquema que propone Vapnik utiliza no sólo funciones de base radial, sino que se proponen otras funciones. En teoría pueden utilizarse todas las funciones simétricas de la forma $K(x, x_i)$ (siendo x_i un vector de soporte) que cumplan el Teorema de Mercer [Vapnik, 1997]. Las funciones de base radial gaussianas no son sino un caso particular. Otras funciones puede que ofrezcan mejor equilibrio entre la tasa de error y la velocidad de convergencia de la red.

- Exploración de otros indicadores de muestra crítica

  En esta tesis sólo se ha examinado un tipo de funciones indicadoras, basadas en un criterio sencillo de selección de muestras. Además, no se ha explorado el comportamiento de las funciones basadas en la norma L_2.

- Entrenamiento de redes con métodos bloque usando muestras críticas

  Las muestras seleccionadas como críticas se pueden utilizar para inicializar la red RBF, y no sólo para construirla. Cuando una muestra nueva es considerada como crítica, puede entrenarse la red en un solo paso con el método LS para reinicializar los pesos. Este método no generalizará bien, aunque es posible que disminuya el riesgo de caída en mínimos locales y acelere la convergencia.

  También puede utilizarse esta misma técnica para entrenar otros esquemas como filtros de Volterra o perceptrones multicapa.

  En general, es interesante probar diferentes tipos de entrenamiento de sistemas con muestras seleccionadas en la forma que se expone en esta tesis.

- Uso de fronteras de referencia no lineales

  En [Lyhyaoui/1, 1999] se utilizan fronteras no lineales, que serán imprescindibles en igualación si el canal es de fase no mínima. Pero además, es posible que en canales de fase mínima su uso reduzca significativamente la tasa de error.

- Selección de varias muestras por centroide

  Se ha sugerido un método para seleccionar varias muestras por centroide. Este método ha sido probado en [Lyhyaoui/1, 1999] con éxito frente a problemas de clasificación en los que una sola muestra es claramente insuficiente. En ese trabajo se prueban cuatro métodos de selección de muestras basados en fronteras de referencia.

  Si se utiliza la frontera RBF formada por los centroides críticos, la función indicadora I_1 tiene un mínimo en torno a la zona de máximo riesgo de la frontera de referencia, pero además tiene otros mínimos que se sitúan en las proximidades de otras zonas de riesgo bayesiano alto. Esto es debido a que la función depende de las posiciones de los centroides, los cuales se sitúan en los máximos de la densidad de probabilidad de los datos.

  Una manera de seleccionar más de una muestra por centroide es buscar estos mínimos. Por desgracia, el problema quizá no tenga solución analítica. Pero es posible que haya soluciones numéricas.

Leganés, Diada de Sant Jordi de 1999

REFERENCIAS

[Adali, 1997] Adali-T; Liu-X; Sonmez-MK, "Conditional distribution learning with neural networks and its application to channel equalization", IEEE Transactions on Signal Processing, vol.45, no.4; April 1997; pp. 1051-1064

[Bernardini, 1993] Bernardini, A., De Fina, S., "A New Predistortion Technique Using Neural Nets", Signal-Processing. vol.34, no.2; Nov. 1993; pp. 231-243

[Biglieri, 1988] Biglieri, E., Barbieris, S., Catena, M., "Analysis and Compensation of Nonlinearities in Digital Transmission Systems", IEEE Journal on Selected Areas in Communications. vol.6, no.1, Jan. 1988, pp. 42-51

[Billings, 1984] Billings, S.A., Gray, J. O., Owens, D.H., Eds., "Nonlinear Systems Design", 1984, Peter Peregrins Ltd. London UK.

[Bishop, 1995] Bishop, C. M., "Neural Networks for Pattern Recognition", Clarendon Press, Oxford, U. K., 1995.

[Burgues, 1997] Burgues, C. J. C., "A Tutorial on Support Vector Machines for Pattern Recognition," in Data Mining and Knowledge Discovery, Fayyad, U., editor, Kluwer Publishers, Boston, November, 1997

[Cachin, 1994] Cachin, C., "Pedagogical Pattern Selection Strategies", Neural Networks, Vol 7, pp. 175-181.

[Cann, 1980] Cann, A. J. "Nonlinearity Model With Variable Knee

Sharpness", IEEE Trans. on Aerospace and Electronic Systems, vol. AES-16, No 6, November, 1980.

[Carini/1, 1995] Carini, Mumolo, "A Novel Algebraic Formulation For The Development of Adaptive Volterra Filtering Algorithms," Proceedings of the NSIP'95, Nonlinear Signal & Image Processing, Thessaloniki, 1995.

[Carini/2, 1995] Carini, Mumolo, "Adaptive Stabilization of Recursive Second Order Polynomial Filters By Means Of a Stability Test," Proceedings of the NSIP'95, Nonlinear Signal & Image Processing, Thessaloniki, 1995.

[Carlson, 1986] Carlson, A. B., "Communication Systems," 3rd ed., McGraw-Hill, 1986

[Cid, 1996] Cid-Sueiro, J., Figueiras-Vidal, A. R., "Channel Equalization with Neural Networks", in "Digital Signal Processing in Telecommunications", Figueiras-Vidal, A. R., Ed., Springer-Verlag, Great Britain, 1996.

[Chen, 1990] Chen, S., Gibson, G. J., Cowan, C. F. N., "Adaptive Equalization of Finite Nonlinear Channels Using Multilayer Perceptrons", Signal Processing, vol 20, No 2, pp. 107-119, 1990.

[Chen, 1991] Chen, S., Gibson, G. J., Cowan, C. F. N., Grant, P. M., "Reconstruction of Binary Signals Using an Adaptive Radial Basis Function Equalizer", Signal Processing, vol. 22, No 2, pp. 77-93, 1991

[Falconer, 1978] Falconer, D. D., "Adaptive Equalization of Channel Nonlinearities in QAM data transmission systems", Bell Syst. Tech. J. Vol. 57, pp 2589-2611., 1978.

[Figueiras/1, 1996] Figueiras-Vidal, A. R., Artés-Rodríguez, A. , Cid-Sueiro, J, Martínez-Ramón, M., "Adaptive Signal Processing: A Discussion Of Trade-Offs from the Perspective of Artificial Learning", Invited Paper, Proc. EUSIPCO, Trieste, Italy, pp. 1635-1640, 1996.

[Figueiras/2, 1996] Figueiras-Vidal, A. R., Martínez-Ramón, M., Artés-Rodríguez, A., Cid-Sueiro, J., "Iterative Decision for Multilievel Equalization", Proceedings of the IASTED Int. Conf. Expert Sys. and NNs, Hawaii, USA., pp. 357-360, 1996

[Figueiras, 1997] Figueiras-Vidal, A. R., Martínez-Ramón, M., "Staircase

Algorithms For Nonlinear Equalization", Proc. COST254: Emerging techniques for communication terminals, pp 175-179, Toulouse, France, 1997.

[Figueiras, 1998] Figueiras Vidal, A. R., Notas del Curso de Doctorado "Tratamiento Estadístico de Señales y Aplicaciones," Universidad Carlos III de Madrid, 1998.

[Fnaiech, 1995] Fnaiech, Guilion, "A Fast M-D Recursive Least Square Adaptive Second Order Volterra Filter," Proceedings of the NSIP'95, Nonlinear Signal & lmage Processing, Thessaloniki, 1995.

[González, 1997] González Serrano, Francisco Javier, "Predistorsión en Comunicaciones Digitales Mediante la Arquitectura CMAC, " Tesis Doctoral, Universidade de Vigo, 1997

[González, 1998] González-Serrano, F. J., Pérez-Cruz, F., Artés-Rodríguez, A., "Reduced Complexity Equalizer for Nonlinear Channels", IEE Electronics Letters, No 9, Vol. 34, pp. 856-858, 1998.

[Goodwin, 1984] Goodwin, Sang Sin, "Adaptive Filtering Prediction and Control," Prentice-Hall, 1984.

[Gradshtein, 1994] Gradshtein, Ryzhik, "Table Of lntegrals, Series and Products,"Academic Press, 1994.

[Graham, 1981] Graham, Alexander, "Kronecker Products and Matrix Calculations With Applications," John Wiley and Sons, New York, 1981

[Haykin, 1991] Haykin, S, "Adaptive Filter Theory," Prentice-Hall, 1991.

[Haykin, 1994] Haykin, S, "Neural Networks-A Comprehensive Foundation", Macmillan, New York, 1994.

[Im, 1995] Im, S., Powers, E.J. "A Third Order Frequency-Domain Adaptive Volterra Filter," Proceedings of the NSIP'95, Nonlinear Signal & lmage Processing, Thessaloniki, 1995.

[Im, 1996] Im, S., Powers, E.J., "A fast method of discrete third-order Volterra filtering", IEEE Transactions on Signal Processing. vol.44, no.9; Sept. 1996; pp. 2195-208

[Isidori, 1989] lsidori, A., "Nonlinear Control Systems," Springer-Verlag, 1989.

[Karam, 1989] Karam, G. Sari, H., "Analysis of Predistortion, Equalization, and ISI Cancellation Techniques in Digital RadioSystems with Nonlinear Transmit Amplifiers", IEEE Trans. on

Comm., vol. 37, No 12, December, 1989.

[Karam, 1991] Karam, G. Sari, H., "A Data Predistortion Technique with Memory for QAM Radio Systems", IEEE Trans. on Comm., vol. 39, No 2, February, 1991.

[Koh, 1985] Koh, T., Powers, E. J., "Second-Order Volterra Filtering and its Application to Nonlinear System Identification", IEEE Transactions on Acoustics, Speech and Signal Processing. vol.ASSP-33, no.6, Dec. 1985, pp. 1445-1455.

[Kohonen, 1989] Kohonen, T., "Self-Organizing and Associative Memory," Springer-Verlag, 1989

[Kohonen, 1990] Kohonen, T., "The Self-Organizing Map," Proc. Of the IEEE, Vol. 78, No. 9, September, 1990, pp. 1464-1477.

[Lihiyaoui/1, 1999] Lihiyaoui, A., Martínez, M., Vázquez, M., Mora, I., Sancho, J., Figueiras, A., "Sample Selection Via Clustering To Construct Support Vector-Like Machines", aceptado para su publicación en IEEE Neural Networks, Special Issue On VC Learning, 1999.

[Lihiyaoui/2, 1999] Lihiyaoui, A., Borrador de Tesis Doctoral. Depositada para su defensa

[Lazzarin, 1994] Lazzarin, G., Pupolin, S., Sarti, A., "nonlinearity compensation in Digital Radio Systems", IEEE Trans. On Communications, vol. 42, No 2/3/4, February/March/April, 1994.

[Lucky, 1975] Lucky, R. W., "Modulation and detection for data transmission on the thelephone channel", New Directions in Signal Processing in Communications and Control, J. K. Skiwirznsky, ed., Leiden, Holland, Noordhoff, 1975.

[Martínez/1, 1999] Martínez-Ramón, M., Sancho-Gómez, J. L., Bousoño-Calzón, C., Figueiras—Vidal, A. R., "Channel Equalization Via Sample Selection", Proc. COST254: Emerging techniques for communication terminals, Neuchâtel, Suiza, 1999. Aceptado para su publicación

[Martínez/2, 1999] Martínez-Ramón, M., González-Serrano, F., Figueiras-Vidal, A. R., "Staircase Equalizers For Nonlinear Channels ", Presentado a IEE Electronics Letters.

[Mathews, 1991] Mathews, "Adaptive Polynomial Filters," IEEE SP Magazine, July 1991, pp. 10-26.

[Minkoff, 1984] Minkoff, J. B., "Wideband Operation Of Nonlinear Solid-State Power Amplifiers-Comparisons of Calculations and Measurements", AT&T Bell Laboratories Technical Journal, Vol 63, No 2 February 1984.

[Mulgrew, 1991] Mulgrew, B., Cowan, C.F.N., "Equalization Techniques Using Non-Linear Adaptive Filters", in Docampo, D., Figueiras, A. R., Eds., "Adaptive Algorithms: Applications and Non Classical Schemes", pp. 1-19, Publicacions da Universidade de Vigo, Vigo, 1991

[Munro, 1992] Munro, P. W., "Repeat Until Bored: A Pattern Selection Strategy", in J. E. Moody, S. J. Hanson, and R. P. Lippmann, eds., Advances in Neural Information Processing Systems, pp. 1001-1008. Morgan Daufmann Publishers, San Mateo, CA., 1992

[Namiki, 1983] Namiki, J., "An Automatically Controlled Predistorter for Multilevel Quadrature Amplitude Modulation", IEEE Transactions on Communications. vol.COM-31, no.5, May 1983, p.707-712

[Nojima, 1980] Nojima, T., Okamoto "Predistortion nonlinear compensator for microwave SSB-AM system", in Conf. Rec., ICC'80, pp. 33.2.1-33.2.6.

[Nowak, 1996] Nowak, R. D., Van Veen, B. D., "Tensor Product Basis Approximations for Volterra Filters," IEEE Transactions on SP, vol.44, no.1; Jan. 1996; pp. 36-50.

[Ochi, 1989] Ochi, Michel K., "Applied Probability and Stochastic Processes In Engineering and Phisical Sciences" John Wiley and Sons, N.y., 1989.

[Oppenheim, 1989] Oppenheim, A. V., Schafer, R., W., "Discrete Time Signal Processing," Prentice Hall, 1989

[Pagés, 1995] Pages-Zamora, A., Lagunas M. A.; Najar, M., Perez-Neira, A., "The K-filter: a new architecture to model and design non-linear systems from Kolmogorov's theorem", Signal Processing. vol.44, no.3; July 1995; pp.249-67.

[Pagés, 1997] Pagés-Zamora, A., Lagunas-Hernández, M. A., "Nonlinear Signal Processing From Fourier Series Based Models", Signal Processing EURASIP, November 1997.

[Powell, 1992] Powell, M. J. D., "The theory of radial basis funcions

approximations in 1990", W. A. Ligth, ed., Advances in Numerical Analysis Volume II, Wavelets, Subdivision Algorithms and Radial Basis Functions, Oxford University, 1992, pp 105-210

[Priestley, 1988] Priestley, "Non-linear and Non-stationary time Series Analysis," Academic Press. 1988.

[Pupolin/1, 1987] Pupolin, S., Greenstein, L. J., "Digital Radio Performance When the Transmitter Spectral Shaping Follows the Power Amplifier", IEEE Trans. on Comm., vol COM-35, No 3, March 1987.

[Pupolin/2, 1987] Pupolin, S., Greenstein, L. J., "Performance Analysis of Digital Radio Links with Nonlinear Transmit Amplifiers", IEEE Journal on Selected Areas in Communications, vol. SAC-5, No 3, April, 1987.

[Saleh, 1983] Saleh, A. A. M., Salz, J. "Adaptive Linearization of Power Amplifiers in Digital Radio Systems", The Bell System Technical Journal, Vol 62, No 4, April 1983.

[Sancho, 1999] Sancho, J., L., Borrador de Tesis Doctoral. Depositada para su defensa.

[Schetzen, 1981] Schetzen, M., "Nonlinear System Modelling Based on The Wiener Theory," Proc. Of the IEEE, Vol 69, No 2 12, 1981.

[Schölkopf, 1997] Schölkopf, B., Sung, K., Burges, C. J. C., Girosi, F., Niyogi, P., Poggio, T., Vapnik, V., "Comparing Support Vector Machines with Gaussian Kernels to Radial Basis Function Classifiers," IEEE Trans. on Signal Proc. Vol 45, No 11, Nov., 1997.

[Sicuranza, 1985] Sicuranza, G. L., "Nonlinear Digital Filter Realization by Distributed Arithmetic," IEEE Transactions on Acoustics, Speech and Signal Processing. vol.ASSP-33, no.4, Aug. 1985, p.939-945

[Sicuranza, 1986] Sicuranza, G. L., Ramponi, G., "Adaptive Nonlinear Digital Filters Using Distributed Arithmetic," IEEE Transactions on Acoustics, Speech and Signal Processing. vol.ASSP-34, no.3, June 1986, pp. 518-526.

[Vapnik, 1982] Vapnik, V., "Estimation of dependences Based on Empirical Data," Nauka, Moscow, 1979 (Springer-Verlag, N. Y., 1982)

[Vapnik, 1995] Vapnik, V. "The Nature of Statistical Learning Theory,"

Springer-Verlag, N. Y., 1995

[Weruaga, 1991] Weruaga-Prieto, L., Cid-Sueiro, J., Figueiras-Vidal, A. R., "Analysis of a look-up table plus separate transversal filter for adaptive nonlinear echo cancellers", First COST 229 WG.2 workshop on Adaptive algorithms, Bayona, Spain, 1991, pp. 142-151.

[Weruaga, 1992] Weruaga-Prieto, L., Cid-Sueiro, J., Figueiras-Vidal, A. R., "Optimal variable-step LMS look-up table plus transversal filter nonlinear echo cancellers", Proceedings of the IEEE ICASSP '92, San Francisco, USA, 1992, vol. 4, pp. 229-232

[Weruaga, 1994] Weruaga-Prieto, L., Figueiras-Vidal, A. R., "Nonlinear Echo Cancelling Using Look-up Tables and Volterra Systems," IEEE Proc.-Vis. Image Signal Processing, vol. 141, no 6, 1994.

[Widrow, 1984] Widrow, B., Walach, E., "On the statistical efficiency of the LMS algorithm with nonstationary inputs," IEEE Trans. On IT, Special Issue on Adaptive Filtering, vol IT-30, No. 2, Pt. 1, 1984, pp. 211-221.

[Widrow, 1990] Widrow, B., Lehr., M. A., "Thirthy years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation," Proceedings of the IEEE, vol. 78, no 9, 1990, pp. 1415-1442.

[Widrow, 1995] Widrow, B., Walach, E., "Adaptive Inverse Control," Prentice-Hall, 1995.

[Wiener, 1958] Wiener, N., "Nonlinear Problems in Random Theory", Wiley and Sons, New York, 1958.

[Yamazaki, 1991] Yamazaki, K., Aly, S., Falconer, D. D., "Convergence behaviour of a jointly-adaptive transversal and memory-based echo canceller", IEEE Proc. F, 1991, 138, (4), pp. 361-370.

[Zhang, 1994] Zhang, B., "Accelerated Learning by Active Example Selection ", International Journal Of Neural Systems, Vol 5, pp. 67-75, 1994

GLOSARIO DE TÉRMINOS

a_n	Símbolo emitido durante el intervalo de tiempo $nT<t<(n+1)T$ (expresión utilizada para tiempo continuo)
B	Ancho de banda en Hz
c	velocidad de la luz en el vacío
c_i	Centroide
c_G	Centroide ganador, o, lo que es lo mismo, más cercano a la muestra presente a la entrada del receptor.
d	Símbolo deseado en recepción. También se utiliza para la distancia entre centroides de clase contraria.
d'	Semidistancia entre muestras críticas más cercanas.
d_i	Distancia del camino i en transmisión con propagación multicamino
f	Variable frecuencial en Hz
F(·)	Modelo de no linealidad a la entrada del canal
G(·)	Modelo de no linealidad a la salida del canal
h(t)	Función de transferencia continua
h[n]	Función de transferencia discreta
h'_k	Atenuación del instante $t=kT$ en la respuesta al impulso del canal en el modelo continuo
h_k	Atenuación del instante $n=k$ en la respuesta al impulso del canal en el modelo discreto
K	Número de elementos de un filtro de Volterra
k_x	Distancia euclídea desde un centroide al plano que corta

	perpendicularmente y en el punto medio del segmento entre este centroide y el más cercano de la clase contraria.
M	Orden del vector de entrada al receptor
$m_x^{b_0 \dots b_j}$	Valor medio de los datos que pertenecen al símbolo que cuyos j primeros bits tienen los valores $b_0 \dots b_j$.
x_j	Símbolo cardinal j ($1 \le j \le P$) del alfabeto en una secuencia PAM de P niveles
N	Orden del canal
n	Variable temporal discreta
n[n]	Vector de ruido de M componentes de entrada al receptor
n[n-k]	Elemento k ($0 \le k \le M-1$) del vector de ruido de entrada al receptor
o(·)	Función de clasificación correspondiente a la salida del filtro central del algoritmo en escalera. También, salida de cualquier clasificador.
$o^{b_0 \dots b_j}$	Función del clasificador del algoritmo en escalera cuyos datos de entrada tienen en común los j primeros bits. La salida determina el valor del bit j+1.
P	Número de elementos del alfabeto o símbolos posibles en una secuencia PAM
P(f), p(t)	Pulso de Nyquist
$P_\beta(f)$, $p_\beta(t)$	Pulso que verifica $P^{1/2}(f)=P_\beta(f)$
r(t)	Vector M componentes de entrada al receptor contaminado con ruido en tiempo continuo
r[n]	Vector M componentes de entrada al receptor contaminado con ruido en tiempo discreto
r[n-k]	Elemento k ($0 \le k \le M-1$) del vector de entrada al receptor más ruido en tiempo discreto
T	Periodo
t	Variable temporal discreta
x[n]	Símbolo emitido en el instante n (expresión utilizada para tiempo discreto)
w	Peso del vector del filtro FIR o de cualquier capa de salida de una red
w	Vector columna de pesos del filtro central del algoritmo en escalera. Vector de pesos de cualquier filtro.
$\mathbf{w}^{b_0 \dots b_j}$	Vector de pesos del clasificador del algoritmo en escalera cuyos datos de entrada tienen en común los j primeros bits. La salida determina el valor del bit j+1.
z	Imagen del x[n] dato en un determinado mapeo.

$\mathbf{z_0}$	Proyección de la imagen del centroide c_i, según un determinado mapeo, sobre la frontera de separación de clases en el espacio que genera el mapeo.
δ	Factor de proporcionalidad de la función de base radial gaussiana que, multiplicado por la distancia al centroide de clase contraria más cercano, da la estimación de la desviación típica de la RBF.
δ_0	Valor teórico del factor de escala.
∇_w	Gradiente de una función respecto de **w**
ε_L	Error calculado como la diferencia entre la señal deseada y la obtenida a la salida del clasificador.
ξ	Amplitud máxima o saturación del canal no lineal.
τ_i	Retardo de propagación del camino i en transmisión con propagación multicamino.
σ_i^2	Varianza o factor de anchura de la RBF pertenenciente al centroide c_i.
σ_x^2	Varianza de x
ω	variable frecuencia angular en rad/s

ABREVIATURAS

AWGN	Additive White Gaussian Noise, Ruido Blanco Gaussiano Aditivo
BP	BackPropagation, Retropropagación
BPSK	Bi-Phase Shift Keying, Modulación por Desplazamiento Binario de Fase
DSP	Digital Signal Processor, Procesador Digital de Señal
FET	Field Effect Transistor, Transistor de Efecto de Campo
FIR	Finite Impulse Response, Respuesta Impulsional Finita
GCMAC	Generalised Cerebellar Model Aritmethic Computer, Calculador Aritmético Modelo Cerebelar Generalizado
ISI	Intersymbol Interference, Interferencia Intersimbólica
IF	Intermediate Frequency, Frecuencia Intermedia
LMS	Least Mean Squares
LUT	Look-Up Table, Tabla de Actualización
MAP	Máximo A Posteriori
MLP	Multilayer Perceptron, Perceptrón Multicapa
MSE	Mean Square Error, Error Cuadrático Medio
PAM	Pulse Amplitude Modulation, Modulación por Amplitud de Pulso
PSK	Phase Shift Keying, Modulación por Desplazamiento de Fase
QAM	Quadrature Amplitude Modulation, Modulación de Amplitud en Cuadratura
QPSK	Quadriphase Shift Keying, Modulación por Desplazamiento de Fase en Cuadratura
RBF	Radial Basis Function, Función de Base Radial
RF	Radiofrecuencia
RLS	Recursive Least Squares, Mínimos Cuadrados Recursivo
SM	Selección de Muestras
SNR	Signal to Noise Ratio, Relación de Señal a Ruido
SVM	Support Vector Machines, Máquinas de Vectores de Soporte
TWT	Travelling Wave Tube, Tubo de Onda Progresiva