Universidad Nacional de La Plata
Facultad de Ciencias Exactas
Departamento de Física

TESIS DOCTORAL

MECANICA ESTADISTICA DEL APRENDIZAJE EN REDES NEURONALES

Luis A. Diambra
Abril de 1996

Departamento de Física

Facultad de Ciencias Exactas

Universidad Nacional de La Plata

Director: Prof. Dr. Angel L. Plastino

A mis padres.

Indice

Sumario

La complejidad de la mente humana ha tenido y tiene fascinados a científicos y filósofos por varios siglos. Desde el siglo XVII, cuando Descartes propuso el aún no dilucidado dilema "mente–cerabro", tanto la Fisiología como la Psicología han debido esperar el aporte de otros campos del conocimiento científico, como el análisis mecánico-estadístico de los modelos neuronales proveniente de la Física, como así también el de las simulaciones con con computadoras, para la mejor comprensión del funcionamiento del cerebro y posiblemente de la mente humana. Las Redes Neuronales (RN) han seducido a los científicos de las más variadas disciplinas, incluyendo neurobiólogos, informáticos, ingenieros y físicos teóricos. Las RN han sido empleadas en modelos de sistemas neurológicos como metáforas para los procesos cognitivos de: aprendizaje, generalización, formación de concepto, etc. También han provisto nuevos modelos de estructuras computacionales y algoritmos para la solución de problemas de optimización y de reconocimiento de patrones. En el terreno de la física teórica, las redes neuronales pueden ser vistas como *nuevos sistemas dinámicos no lineales de nodos interconectados* con una amplia gama de comportamientos. Desde el punto de vista de la Mecánica Estadística, estos sistemas exhiben diversos grados de desorden y frustración. Por esta razones, el campo de RN es de creciente interés y registra intensa actividad científica y tecnológica, aplicable a diversos areas de la ciencia.

En este trabajo de Tesis Doctoral, nos dedicaremos en particular a investigar problemas relativos a los algoritmos de aprendizaje, con hincapié en arquitecturas de redes tipo *feedforward.* Es de especial interés en este ámbito determinar bajo qué condiciones una red es capaz de captar la regla general subyacente en los ejemplos que se le muestran. Esta habilidad es conocida como generalización y ha sido estudiada

en la literatura especializada, en el marco de la Mecánica Estadística, encontrándose transiciones de fase entre estados termodinámicos que incluyen fases de generalización perfecta, fases de generalización pobre, y fases de vidrios de spin donde el estado fundamental está fuertemente degenerado y el sistema se encuentra *congelado*.

El cuerpo principal de esta Tesis está organizado en cuatro capítulos, de la siguiente manera:

En el Capítulo 1 repasamos brevemente los conceptos básicos relacionados con el sistema nervioso y especialmente la neurona, que nos permitirán construir modelos de redes nerviosas. También introducimos los rudimentos de la teoría del aprendizaje. En el Capítulo 2 investigaremos, con herramientas de la Mecánica Estadística y el método de Réplica, en qué condiciones el perceptrón es capaz de aprender una regla general a partir de cierto conjunto de ejemplos. En este capítulo, el proceso de aprendizaje es entendido como un proceso estocástico de relajación. Obviamente, existen muchas formas de lograr que una RN aprenda una tarea dada, y uno debe considerar criteriosamente un compromiso apropiado entre la regla a aprender, el número de ejemplos disponibles y el costo computacional. En el Capítulo 3 introducimos un nuevo formalismo para el aprendizaje, basado en el Principio de Máxima Entropía, en el marco de la Teoría de la Información. En este esquema, el proceso de aprendizaje es considerado como un proceso de inferencia, y no como una relajación. Por último, en el Capítulo 4, mostramos algunos ejemplos importantes de aplicación de nuestros métodos computacionales. Dos problemas Físicos son atacados: la predicción de series temporales caóticas y la reconstrucción de funciones de onda a partir de información parcial.

Capítulo 1

Introducción al modelado en Redes Neuronales

1.1 El sistema nervioso y su unidad

Una de las primeras apreciaciones que, desde un punta físico, debiera tenerse en cuenta para entender el comportamiento del cerebro (como el de muchos otros sistemas complejos), es la distinción entre dos planos descriptivos. Desde un punto de vista *reduccionista*, el cerebro como porción de materia ordinaria tiene las propiedades comúnmente asociadas con ésta: temperatura, potenciales químicos, campos eléctricos, etc. Por otro lado, desde un punto de vista más integrado, *holista,* el cerebro exhibe una nueva clase de fenómenos que no son observados en los niveles más bajos de organización de la materia. Estas son las propiedades colectivas emergentes, asociadas con el nivel cognitivo: almacenamiento y recuperación de memoria, asociación, reconocimiento de patrones, categorización, generalización, aprendizaje, etc. Son estas propiedades las que las ciencias cognitivas pretenden comprender con el lenguaje de la física de los sistemas complejos, modelando la red nerviosa de los organismos vivos como un sistema de elementos interactuantes cuya evolución es gobernada por leyes dinámicas definidas.

El cerebro es un sistema complejo por excelencia. Está formado por un número del orden de 10^{11} células nerviosas, llamadas *neuronas*, las que están conectadas entre

sí por alrededor de 10^{15} junturas sinápticas [1, 2]. Una única célula puede estar conectada con otras 10000 células, vecinas o no, por medio de hasta 200000 junturas sinápticas. Los patrones de conectividad son extramadamente intrincados y varían de acuerdo con las funciones de las células dentro del sistema nervioso.

Existe un gran variedad de tipos de células nerviosas que pueden ser divididas en diferentes categorías según el criterio adoptado [1, 2]. Sin embargo, para nuestro propósito es suficiente considerar el esquema más simple. La arquitectura de las células nerviosas consta básicamente de las mismas partes, sin importar su tamaño o forma. El cuerpo celular o *soma*, es el responsable de la actividad metabólica de la célula. Desde el soma se proyectan diferentes ramificaciones, las *dendritas*, como así también una fibra tubular, llamada *axón*, que se bifurca al final en numerosas ramas (ver Fig. 1.1). Las dendritas actúan como receptoras de las señales provenientes de otras células. La información contenida en estas señales es procesada o *integrada* en el soma para dar una respuesta de salida que es transmitida por el axón, en una forma muy eficiente, hacia las junturas sinápticas que afectan a las neuronas postsinápticas o a una fibra muscular. Las *sinapsis*, son las junturas entre una terminal axónica de una neurona presináptica (eferente) con una dendrita o con el soma de una neurona postsináptica (aferente).

Las señales nerviosas son transmitidas eléctrica o químicamente. La transmisión eléctrica prevalece en el interior de una neurona, mientras que los mecanismos químicos de transimisión operan entre neuronas diferentes a través de las sinapsis. Para los modelos de RN, sólo son de interés estas últimas, que discutiremos brevemente. Una vez que la señal arriba a la terminal de la neurona presináptica, ésta libera pequeñas cantidades de uno o más neurotransmisores químicos (NTQ) desde unas vesículas ubicadas en las terminales del axón. Estas móleculas difunden a través de cierto espacio sináptico hasta los sitios receptores (ver Fig. 1.1) ubicados en la membrana postsináptica, afectando su permeabilidad con respecto a ciertos iones (mayormente Na^+y K^+) [1]. En el estado de inactividad, el interior de la célula está cargado negativamente con respecto al medio intercelular. Este potencial de reposo, alrededor de -70 mV, se debe a las diferentes concentraciones de iones entre la célula y el medio. Los gradientes de concentraciones son mantenidos por bombas (proteínas) en

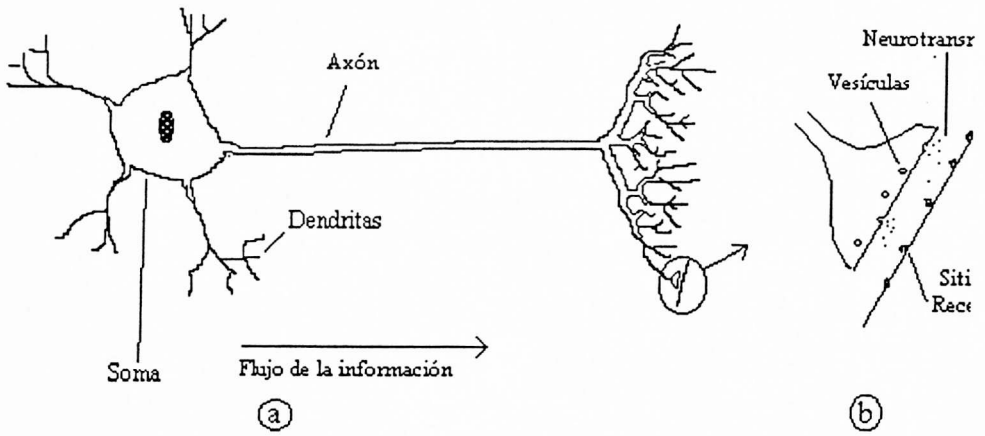

Figura 1.1: Diagrama simplificado de una neurona típica (a). Esquema de una juntura sináptica (b).

la pared celular, que son impenetrables para los iones de Na^+. Por efecto de los NTQ, la permeabilidad de la membrana cambia, y se altera el potencial de membrana, que es la diferencia de potencial entre la célula y el medio. Después de este cambio en potencial de menbrana, existe cierto período, llamado período refractario, durante el cual la neurona es incapaz de responder a otros estímulos. Las fluctuaciones del potencial de membrana pueden ser positivas (correspondientes a un estímulo excitatorio) o negativas (estímulos inhibitorios), dependiendo de la naturaleza de los NTQ y de los sitios receptores.

Además de mantener los procesos metabólicos comunes a toda célula, el soma está especializado en la integración de los impulsos provenientes de diferentes neuronas presinápticas. Estos impulsos deben ser pesados en forma diferente, debido a la variedad de propiedades intrínsecas que presentan las dendritas y sus diferentes localizaciones con respecto al soma. En este sentido, se habla de eficacias o *pesos sinápticos*. El efecto neto del proceso de integración (que puede ser no lineal [3]) es producir un potencial que al superar cierto *umbral de activación* es transmitido por el axón a otras células; decimos entonces que la neurona "dispara".

Existe una enorme cantidad de evidencia acerca de que los pesos sinápticos, no

son fijos. Como originalmente postulara Donald Hebb [4], los pesos sinápticos pueden ser ajustados si su nivel de actividad cambia; así el peso sináptico de una neurona que repetidamente dispara, (alto nivel de actividad) es fortalecido, mientras que otros son debilitados. Este mecanismo de *plasticidad* sináptica juega un rol fundamental en los complejos procesos de aprendizaje.

Existen muchas otras propiedades o características biológicas de las neuronas, y del sistema nervioso en general, a ser tenidas en cuenta en modelos sofisticados de funciones cerebrales más específicas [3, 5]. Sin embargo, en este momento nos hallamos con la colección mínima de hechos relevantes que nos permiten elaborar algunos modelos de RN, en particular los que son el tema de estudio en este trabajo de Tesis.

1.2 Modelos de Redes Neuronales

Las características biológicas a ser tenidas en cuenta aquí para la construción de modelos de RN fueron consideradas primeramente por Warren McCulloch y Walter Pitts [6], quienes establecieron las operaciones lógicas para una neurona *formal.* Las versiones algebraicas de este modelo fueron estudiadas entre los años '60 y '70 [7, 8, 9]. Las principales características consideradas son: el nivel de actividad o estado de una neurona, el potencial de membrana, el umbral de activación de la neurona y los pesos o eficacias sinápticas, como así también algún mecanismo de plasticidad sináptica para el aprendizaje. En términos matemáticos, un modelo de red neuronal está definido como un grafo orientado, con las siguientes propiedades:

- Una variable de estado n_i está asociada con cada nodo (neurona) i.
- Un peso (peso sináptico) w_{ij} está asociado con cada par ordenado (i, j) de nodos.
- Cada nodo i tiene asociado un umbral (umbral de activación) θ_i.
- Una función de transferencia f_i, definida para cada nodo i, determina el estado de un nodo como una función del umbral, de los pesos sinápticos y de los estados de los nodos conectados.

La función de transferencia establece la ley dinámica con la cual evoluciona la red y usualmente es la misma para todas las neuronas con el fin de facilitar el análisis estadístico. La forma es $f(h_i - \theta_i)$, donde $h_i = \sum_j w_{ij} n_j$ es el potencial de membrana de la neurona i, debido a la acción de las otras neuronas[1]. La función f es o bien una función escalón (como en los modelos con estados discretos), o bien una función *sigmoidal.* Los pesos sinápticos describen el acoplamiento entre dos neuronas. En cierto sentido, el *programa* que gobierna el proceso computacional está dado por la matriz de acoplamientos w_{ij}, que en general puede ser no simétrica.

Otra importante cuestión a discutir es el punto relativo al conexionado o *arquitectura* de la red. Existen dos arquitecturas básicas y muchas otras que interpolan entre ambos extremos. La arquitectura *fully-connected* es aquella donde todas las neuronas están acopladas entre sí. Está arquitectura esta profundamente ligada a los sistemas magnéticos y provee un sustrato para el almacenamiento y recuperación de memorias, para problemas de optimización, etc.. Por otro lado, emulando los tejidos vivos, existe una arquitectura de redes a capas, conocida como *feedforward.* En esta arquitectura, las neuronas de una capa sólo afectan a las neuronas de la siguiente capa; es decir, el flujo de la información es unidireccional. Dichas arquitecturas están asociadas mayormente al reconocimiento de patrones, categorización y generalización, pero pueden ser usadas también como memoria asociativa.

Analicemos brevemente la primera de las arquitecturas mencionadas, la *fully-connected* (Fig. 1.2). En dicho caso, las reglas para la actualización (ley dinámica) de los estados del sistema puede ser determinística o probabilística. En cuanto al momento en que se hace efectiva dicha actualización, las reglas pueden ser sincrónicas o asincrónicas. El modelo inicial de McCulloch–Pitts [6] considera neuronas con sólo dos estados, es decir que la variable de estado n_i puede tomar los valores 1 si la neurona está activa, ó -1 si está en reposo. Los cambios en los estados de la red ocurren en pasos de tiempo discreto $t = 0, 1, 2, ...$, en forma sincrónica, o sea que todas las neuronas son actualizadas al mismo tiempo. El estado de la neurona i al

[1] El potencial de menbrana puede ser una función complicada de la actividad de las otras neuronas y muchas veces es necesario incorporar cierto grado de no linealidad en el proceso de integración.

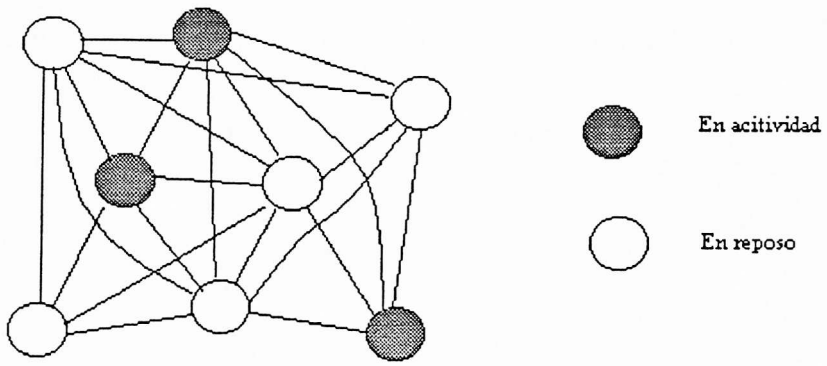

Figura 1.2: Esquema de una red *fully-connected* con neuronas binarias.

tiempo t, está dado por la siguiente ley de evolución dinámica determinista:

$$n_i(t) = \text{sgn}\left[h_i(t-1) - \theta_i\right] \tag{1.1}$$

donde sgn(x) es la función signo y h_i el potencial de membrana.

Existen algunas extensiones de este modelo, introducidos por William Little [9] y por John Hopfield [10, 11], en los que se incorpora una ley dinámica estocástica (modelos no determinista.) En este caso, la regla de actualización (1.1), es reemplazada por:

$$P(n_i(t) = 1) = \frac{1}{1 + \exp\left[-\beta\left(h_i(t-1) + \theta_i\right)\right]} \tag{1.2}$$

donde β es una medida de aleatoriedad. La ecuación (1.2) sólo da la probabilidad de que la i-ésima neurona tome el valor $+1$. Tomando el límite $\beta \to \infty$, recuperamos el modelo determinista.

La diferencia entre el modelo de Little y el de Hopfield está en la forma en que el sistema es actualizado. En el primero, todas las neuronas son actualizadas sincrónicamente. Mientras tanto en el modelo de Hopfield el estado de las neuronas es actualizado secuencialmente (dinámica de Glauber [12]), ya sea con algún orden o en forma aleatoria. La simulación secuencial tiene ventajas considerables a la hora de simular la red en una computadora convencional y para el análisis teórico de las

propiedades de la red; pero en este caso dejamos de lado las ventajas operacionales del procesamiento en paralelo, característica básica de las redes neuronales. Las neuronas de nuestro cerebro no operan secuencialmente, siendo ésta la razón de la superioridad del cerebro, en tareas complejas, frente a las computadoras electrónicas más rápidas.

Por otro lado, existe una fructífera analogía entre estos modelos de redes con pesos sinápticos simétricos y los sistemas de spines (modelo de Ising [13]). Cada spin puede estar en dos estados *up* y *down*, correspondientes a neurona activa ($n = 1$) y en reposo ($n = -1$), respectivamente. A partir de esta analogía, podemos explotar las propiedades físicas mejor conocidas de los sistemas magnéticos [14], a fin de comprender algunas de las propiedades de los modelos que aquí nos interesan.

En términos de un sistema de Ising, un spin interactúa con otro, induciendo un campo magnético. El campo magnético total sobre el spin i debido a los otros spines está dado por $h_i = \sum_j w_{ij}\, n_j - \theta_i$, siendo θ_i el campo externo sobre la neurona i. Si el sistema consta de N de tales spines, entonces existen 2^N configuraciones con una energía definida por

$$E[n] = -\sum_{ij=1}^{N} w_{ij}\, n_i\, n_j + \sum_{i=1}^{N} \theta_i\, n_i \tag{1.3}$$

donde los términos de autoenergía w_{ii} no afectan a la dinámica y pueden ser excluídos de la suma. Si los w_{ij} son todos positivos, el material es ferromagnético, en cambio si los acoplamientos tienen signos y valores absolutos distribuídos al azar, el material es un vidrio de spin [14, 15, 16]. Dentro de esta analogía el parámetro β introducido en (1.2) corresponde con la inversa de la temperatura del baño térmico.

Una diferencia sustancial de los modelos de RN con los sistemas magnéticos, consiste en la naturaleza de los acoplamientos sinápticos entre las neuronas. Estos acoplamientos son en general no simétricos, es decir que $w_{ij} \neq w_{ji}$. Esta violación de la Tercera Ley de Newton en el caso de los tejidos vivos indica que no es posible definir una función energía que siempre decrezca bajo la dinámica. Por fortuna, esto no afecta a las características esenciales de almacenamiento y evocación de patrones (memorias) que presentan estos sistemas.

La memoria puede ser definida como el estado particular de un sistema, cuyo comportamiento ha sido alterado por determinado estímulo. Una memoria asociativa

almacena y recupera información por asociación con otra información o con conocimiento parcial de su contenido. Dada una red con sus pesos sinápticos y umbrales definidos tiene ciertos conjunto de puntos fijos o ciclos límites (atractores del sistema), que pueden ser considerados como memorias. Cada memoria consistiría entonces en un pequeño subconjunto de los 2^N configuraciones, y el estímulo estaría dado por la condición inicial impuesta a la red como información parcial. En virtud de su dinámica, la red se mueve hacia el atractor y alcanza el ciclo límite asociado con la memoria almacenada. El problema de cómo *sintonizar* los parámetros de la red (pesos sinápticos, umbrales de activación, etc.) para que determinadas configuraciones sean atractores del sistema, constituye la tarea central del proceso de apredizaje, el que será tratado con detalle en capítulos siguientes.

1.3 Redes cibernéticas

Los modelos de RN, ampliamente discutidos a partir de los trabajos de Hopfield [10, 11], están relacionados con los sistemas de almacenamiento y recuperación de memorias. La memoria es una importante función del cerebro pero no es la única. Otra de las funciones del sistema nervioso central consiste en el aprendizaje de las reacciones y comportamientos que le permitirán al individuo sobrevivir en un medio ambiente hostil. Detectar una señal en el *ruido* originado por el entorno, puede ser la diferencia entre sobrevivir o morir. Sin dudas, ésta debió ser la *fuerza* que originalmente empujó el desarrollo del cerebro en nuestro proceso evolutivo.

Introducimos la expresión redes cibernéticas para denominar aquellos circuitos neuronales capaces de proveer una reacción satisfactoria a un estímulo externo. Tales redes tienen una estructura que difiere radicalmente de las redes que almacenan memorias, puesto que sus pesos sinápticos son fuertemente asimétricos o incluso unidireccionales. Como resultado de esta asimetría, la teoría termodinámica de sistemas en equilibrio no es aplicable directamente a las redes cibernéticas, al menos en lo concerniente a su comportamiento dinámico.

La clase de redes cibernéticas mejor estudiadas son las conocidas como redes de capas *feed-forward*, donde el flujo de la información es en una única dirección entre

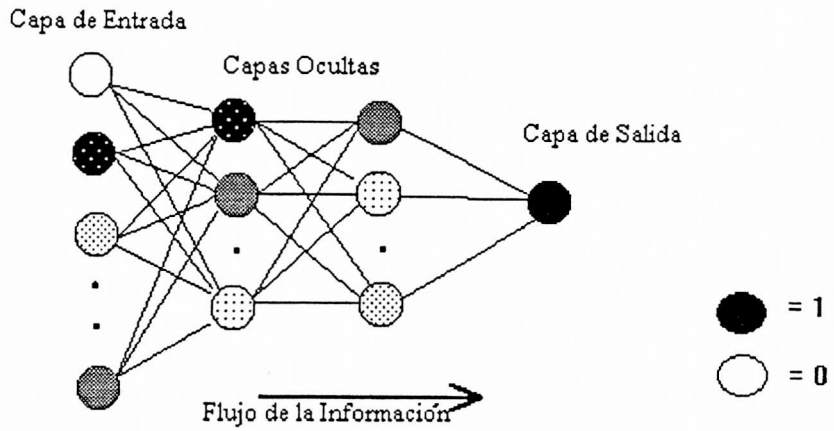

Figura 1.3: Esquema de una red *feed-forward* de varias capas.

las distintas capas de neuronas (ver Fig. 1.3). La primera capa (capa de entrada), está compuesta por neuronas sensoras que reciben el estímulo externo; las capas intermedias lo procesan; y finalmente las neurona de la capa de salida, producen la respuesta que puede ser derivada a otra red para un procesamiento ulterior o bien puede derivar en un señal a las neuronas motoras.

En términos matemáticos, un cierto conjunto de relaciones de *entrada–salida* llamado conjunto de entrenamiento definen un mapeo, y la red *feed-forward* provee una representación de este mapeo. El problema más importante vinculado a estos sistemas, a ser tratados aquí, consiste en cómo podemos establecer una dada representación del mapeo subyacente a los pares entrada–salida, por medio de un algoritmo o regla de aprendizaje. El primero en estudiar este problema fue Rosemblatt [17], quien introdujo el modelo más sencillo de red *feed-forward*, conocido como *perceptrón.* En este modelo, las N neuronas de entrada están directamente conectadas a las de salida. Denotaremos el estado de las neuronas de entrada por S_i $(i = 1, 2, ..., N)$, mientras que las neuronas de salida serán denotadas por la variable ξ_i. El estado de la i-ésima neurona de salida está determinado por

$$\xi_i = f\left(\sum_{j=1}^{N} w_{ij}\, S_j - \theta\right), \tag{1.4}$$

donde θ es el umbral de activación. La función f puede ser considerada como una ley estocástica la cual determina la probabilidad de que $S_i = \pm 1$, o bien como una salida continua. Aquí nos concentraremos en este último caso.

Para cada matriz de pesos w_{ij}, la red realiza un mapeo de la entrada, denotada por $\mathbf{S}$, sobre la salida ξ. El entrenamiento del perceptrón consiste en una adecuada elección de los pesos w_{ij} sobre la base de la información contenida en P ejemplos dados por los pares entradas–salidas $\{\mathbf{S}^\mu, \xi^\mu\}$ con $\mu = 1, 2, ..., P$, de forma tal que el perceptrón asocie a cada entrada $\mathbf{S}^\mu$ la correspondiente salida ξ^μ. No existe una función explícita que permita calcular, en forma general, los pesos de la red a partir de los ejemplos. Sin embargo, es posible construir algoritmos iterativos, que pueden converger a los valores deseados de los pesos sinápticos si éstos existen. Señalemos además que cuando la red no es capaz de aprender la regla general, aún con una gran cantidad de ejemplos, se dice que el problema es noaprendible. La razón de este hecho puede encontrarse en que no existe un conjunto de pesos que mapee correctamente todos los ejemplos. En ese caso, la arquitectura de la red no es la adecuada para realizar esa tarea.

Hebb [4] enunció su ley de plasticidad sináptica: "Cuando una neurona presináptica j excita repetidamente a una neurona postsináptica i, tiene lugar en las neuronas un proceso tal que la eficacia sináptica w_{ij} aumenta". Esta ley constituye una base para la inspiración de los algoritmos de aprendizaje en redes computacionales artificiales. Para su implementación, consideremos el caso simple donde las neuronas de salida son binarias y solo un patrón de comportamiento (un único ejemplo) es aprendido. En este caso, $\xi_i = \xi_i^1$ si la reacción de la neurona i es la correcta o bien $\xi_i = -\xi_i^1$ si es incorrecta. Comenzamos con una configuración inicial de pesos al azar, el peso sináptico es incrementado en la cantidad:

$$\begin{aligned} \delta w_{ij} &= \varepsilon \left(1 - \xi_i \xi_i^1\right) \xi_i^1 S_j^1 \\ &= \varepsilon \left(\xi_i^1 - \xi_i\right) S_j^1 \end{aligned} \tag{1.5}$$

luego calculamos las nuevas salidas ξ_i y repetimos el paso (1.5). A partir de el último término de (1.5) podemos interpretar que si la reacción es correcta, el peso no se modifica, mientras que si es incorrecta el peso es modificado. El parámetro

ε caracteriza la intensidad del proceso de aprendizaje, y debe ser elegido en forma tal que el algoritmo converja en un tiempo razonable, cuando aplicamos repetidas veces el proceso. En el caso en que varios patrones deben ser aprendidos, podemos generalizar la expresión (1.5) como

$$\delta w_{ij} = \varepsilon \sum_{\mu} (\xi_i^{\mu} - \xi_i) \, S_j^{\mu} \tag{1.6}$$

Si bien este mecanismo tiene cierto correlato biológico y puede ser de utilidad para estudiar el comportamiento de seres vivos simples; existen otros mecanismos, sin correlato biológico, utilizados para entrenar redes artificiales. En este sentido, la forma más usual es considerar el proceso de aprendizaje como un proceso de minimización de cierta función costo o energía. Existen diversos protocolos para la minimización de una función: gradiente descendente [18], simulación de templado [19] o algoritmos genéticos [20]. Estas rutinas pueden incorporar o no cierto grado de aleatoriedad, representado ya sea por una temperatura, o bien por las mutaciones, y evitar asíque el sistema quede atrapado en una mínimos relativos de baja performance. Las ventajas de estas formas de búsqueda de mínimos, es que pueden ser extendidas para el entrenamiento de redes con más capas, útiles en problemas de gran complejidad.

Existen dos cuestiones muy importantes relativas al aprendizaje. La primera consiste en cómo evaluar la *performance* de una red ya entrenada para realizar una dada tarea, y la otra, en qué condiciones la red es capaz de mapear correctamente una entrada que no estaba en el conjunto de entrenamiento. Esta habilidad es conocida como capacidad de generalización. En el próximo capítulo estudiaremos una forma de cuantificar dichas propiedades.

Bibliografía

[1] G.M. Shepherd, *The Organization Sinaptic of the Brain* (Oxford University Press, Oxford, 1979, 2^{da} edición).

[2] E.R. Kandel, *Cellular Basis of Behavior* (W.H. Freeman, San Fransisco, 1976).

[3] J.W. Clark, *Introduction to Neural Networks, in Nolinear Phenomena in Complex Systems*, Editado por A.N. Proto (Elsevier, Amsterdam, 1989).

[4] D.O. Hebb, *The Organization of Behavior: A Neuropsychological Theory* (Wiley, New York, 1949).

[5] D. Amit, *Modelling Brain Functions* (Cambridge University, Cambridge, 1989).

[6] W.S. McCulloch y W. Pitts, Bull. Math. Biophys. **5**, 115 (1943).

[7] E.R. Caianiello, J. Theoret. Biol. **2**, 204 (1961).

[8] E.M. Hart, T.J. Csermely, B. Beek, y R.D. Lindsay, J. Theoret. Biol. **26**, 93 (1970).

[9] W.A. Little, Math. Biosci. **19**, 101 (1974).

[10] J.J. Hopfield, Proc. Nat. Acad. Sci. **79**, 2554 (1982).

[11] J.J. Hopfield y D.W.Tank, Science **241**, 625 (1986).

[12] R.J. Glauber, J. Math. Phys. **4**, 294 (1963).

[13] E. Ising, Z. Physik **31**, 253 (1925).

[14] S.F. Edwards y P.W. Anderson, J. Phys. F **5**, 965 (1975).

[15] D.J. Amit, H. Gutfreund, y H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985).

[16] D.J. Amit, H. Gutfreund, y H. Sompolinsky, Ann. Phys. **173**, 30 (1987).

[17] F. Rosemblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanism* (Spartan Books, Washington, D.C., 1962).

[18] Y. Le Cun, *Disordered Systems and Biological Organizations*, editado por E. Bienenstock, F. Fogelman, y G. Weisbuch (Springer, Berlin, 1986).

[19] S. Kirkpatrick, C. Gellat, y M. Vecci, Science **220**, 671 (1983).

[20] J. Holland, *Evolution, Learning and Cognition,* editado por Y.S. Lee (World Scientific, Singapore, 1988).

Capítulo 2

Termodinámica del aprendizaje

2.1 Introducción

Como es natural, para que una red neuronal realice una tarea determinada, es menester llevar a cabo una etapa previa de aprendizaje o entrenamiento. Tanto el entrenamiento, como la posterior evaluación de la red, dependen de la tarea a realizar. En particular, en las tareas que involucran almacenamiento y recuperación de memoria, el entrenamiento de la red consiste en encontrar los pesos sinápticos tal que los patrones sean puntos fijos de la dinámica del sistema neuronal. La etapa de evaluación de la performance consiste básicamente en testear la capacidad de reconstrucción de los patrones almacenados, cuando presentamos como condición inicial de la dinámica de la red los patrones levemente distorsionados. Este tipo de memoria se conoce como memoria asociativa.

Sin embargo, una red neuronal debidamente entrenada, también es capaz de asociar patrones totalmente diferentes. En este caso, puede existir o no algún tipo de correlación entre el conjunto de patrones iniciales (entradas) y el conjunto de patrones finales (salida). Si no existe dicha correlación estamos frente a una memoria heteroasociativa. En cambio, si los patrones consisten en pares entrada–salida, que manifiestan un mapeo específico, es decir, que existe una función que asigna a cada entrada una salida, estamos ante un problema diferente, pues existe una regla general en el conjunto de ejemplos ó conjunto de entrenamiento (CE). Es posible que la red

no solo memorice los patrones del CE, sino que además logre captar la regla subyacente en los ejemplos pertenecientes al CE, es decir aprenda a mapear correctamente una señal de entrada, aún cuando esta no es miembro del CE. Esta habilidad de las redes se conoce como *generalización* y constituye una propiedad muy importante de las redes neuronales, en especial las de arquitecturas *feedforward* [1].

Cuando pretendemos que una red neuronal aprenda una dada regla, tenemos que buscar un conjunto de parámetros, pesos sinápticos y umbrales de activación, que otorgen a la red neuronal la capacidad no solo de asociar correctamente los patrones de CE, sino que además pueda generalizar. Estos son elementos para tener en cuenta en la elección del método de entrenamiento, que la red debe asociar correctamente con su salida, cualquier señal de entrada y no solamente las pertenecientes al CE. Dentro del marco del aprendizaje supervisado [1, 2, 3, 4, 5, 6, 7], el CE puede ser generado por otro perceptrón con peso $\mathbf{W}_0$, llamado perceptrón maestro (PM). En esta instancia, se presentan dos tipos de situaciones: reglas aprendibles y reglas no aprendibles [8]. Las reglas no aprendibles están vinculadas a la diferencia en la arquitectura (es decir, función de transferencia y espacio de pesos) entre el PM y el perceptrón a ser entrenado, llamado perceptrón estudiante (PE). Si tanto PM como PE tienen la misma arquitectura, la regla es aprendible pues existe al menos un vector en el espacio de los pesos que puede aprender en forma exacta la regla subyacente a los ejemplos.

En este capítulo, investigaremos el aprendizaje de una regla a partir de ejemplos en el tipo más simple de redes de capas *feedforward:* el perceptrón de una sola capa [1, 9], con las herramientas derivadas de la Mecánica Estadística y el método de Réplica [4, 10, 11, 12], entendiendo el aprendizaje como un proceso de relajación. Primeramente, estudiaremos la existencia de transiciones de fase a estados con generalización perfecta, en el límite de altas temperaturas. En el marco de una aproximación perturbativa, incorporamos el desorden provenientes de la aleatoriedad de los ejemplos, mediante la consideración de las interacciones entre dos réplicas. Este formalismo nos permite estudiar con bastante exactitud la termodinánica del perceptrón, en el rango de temperaturas donde existe simetría de réplica. Por último, focalizaremos nuestra atención en estudiar, con esta nueva herramienta, la posibilidad que un perceptrón logre generalizar cuando el CE está contaminado con ejemplos erroneos, es decir que

no son generados por el PM.

2.2 Formalismo general

Dada una red que consiste en N unidades de entrada S_i conectadas a una unidad de salida ξ, cuyo estado es determinado por

$$\xi = g\left(N^{-1/2}\mathbf{W}\cdot\mathbf{S}\right), \tag{2.1}$$

donde $g(x)$ es la función de transferencia, de forma tal que para cada vector de pesos $\mathbf{W} \equiv (W_1, ..., W_N)$ la red realiza un mapeo de $\mathbf{S} \equiv (S_1, ..., S_N)$ a ξ. El proceso de aprendizaje tiene lugar cuando los pesos W_i entre las unidades de entrada y la salida, son sintonizados en forma tal que la red se acerque al mapeo deseado $\xi_0 = g\left(N^{-1/2}\mathbf{W}_0\cdot\mathbf{S}\right)$, tanto como sea posible. Una de las formas de alcanzar este objetivo, es a travez de minimizar una cierta de función costo o energía de entrenamiento E_t, construida a partir de un CE con P ejemplos $\{\mathbf{S}^l, \xi_0(\mathbf{S}^l)\}$ con $l = 1, ..., P$. En este caso, la función energía de entrenamiento E_t a minimizar, está definida por

$$E_t(\mathbf{W}) = \sum_{l=1}^{P} \epsilon\left(\mathbf{W}, \mathbf{S}^l\right), \tag{2.2}$$

donde $\epsilon(\mathbf{W}, \mathbf{S})$ recibe el nombre de *función error*, y expresa una medida de la desviación entre la salida de la red $\xi(\mathbf{W}, \mathbf{S})$ y la salida deseada $\xi_0(\mathbf{S})$. La performance de la red sobre todo el espacio de los ejemplos es medido por la *función de generalización* $\varepsilon(\mathbf{W})$, la cual está definida, como el promedio de la función error sobre todo el espacio de las entradas

$$\varepsilon(\mathbf{W}) = \int d\mu(\mathbf{S})\, \epsilon(\mathbf{W}, \mathbf{S}), \tag{2.3}$$

donde $d\mu(\mathbf{S})$ denota una medida. En este capítulo, estudiaremos el aprendizaje como un proceso de relajación estocástico, en el cual los pesos de la red evolucionan de acuerdo con la ecuación de Langevin

$$\frac{\partial \mathbf{W}}{\partial t} = -\nabla_w E_t(\mathbf{W}) + \eta(t), \tag{2.4}$$

donde η es ruido gaussiano con media nula y varianza:

$$\langle \eta_i(t)\,\eta_j(t')\rangle = 2T\delta_{ij}\delta(t-t'). \tag{2.5}$$

Esta dinámica tiende a disminuir la energía E_t, pero ocasionalmente la energía puede aumentar como consecuencia de las fluctuaciones térmicas. A $T=0$, las fluctuaciones desaparecen y la ecuación (2.4) corresponde simplemente a un gradiente descendente. Así si la superficie de energía es suave[1], el sistema alcanza rápidamente el estado fundamental, correspondiente al mínimo global de su energía. La justificación de la incorporación del baño térmico en la dinámica, la podemos encontrar en el hecho que la superficie de energía es por lo general muy rugosa, con una enorme cantidad de máximos y mínimos, por lo que el sistema puede quedar atrapado en alguno de sus mínimos locales, no deseados por su baja performance. Las fluctuaciones térmicas proveen un mecanismos para escaparse de esos mínimos locales.

Sabemos que a tiempos suficientemente grandes, la dinámica (2.4) conduce a una distribución de probabilidades tipo Gibbs para los pesos de la red [13]

$$P(\mathbf{W}) = Z^{-1}\exp[-\beta E_t(\mathbf{W})], \tag{2.6}$$

donde $\beta = 1/T$ y T es la temperatura y denota el nivel de ruido en el proceso de aprendizaje. El factor de normalización Z es la función de partición

$$Z = \int d\mu(\mathbf{W})\exp[-\beta E_t(\mathbf{W})]. \tag{2.7}$$

Como la energía depende de la elección particular los ejemplos, el sistema presenta un nuevo tipo de desorden a ser tenido en cuenta: el desorden *quenched*. Por esta razón, los observables macroscópicos del sistema deben ser calculados promediando sobre dos espacio: un promedio térmico sobre el espacio de los pesos con distribución $P(\mathbf{W})$ y es denotado por $\langle ... \rangle_T$, y el promedio sobre el espacio de los ejemplos, llamado promedio *quenched*, que sera denotado por $\ll ... \gg \equiv \int ... \prod_l d\mu(\mathbf{S}^l)$.

La energía libre F de la red está dada por

$$F(T,P) = -T \ll \ln Z \gg, \tag{2.8}$$

$$S(T,P) = -\ll \int d\mu(\mathbf{W})\,P(\mathbf{W})\ln P(\mathbf{W}) \gg . \tag{2.9}$$

[1] Es decir, con un único mínimo.

La performance de la red con respecto a los ejemplos que no pertenecen al CE, es medido por el error de generalización ϵ_g, mientras que la bondad de la red con respecto al CE, es el error promedio de entrenamiento ϵ_t

$$\epsilon_t(T,P) = P^{-1} \ll \langle E(\mathbf{W})\rangle_T \gg, \tag{2.10}$$

$$\epsilon_g(T,P) = \ll \langle \varepsilon(\mathbf{W})\rangle_T \gg . \tag{2.11}$$

Las cantidades (2.8),(2.9) y (2.10) están relacionadas por la identidad

$$F = P\epsilon_t - TS. \tag{2.12}$$

Las gráficas de $\epsilon_t(T,P)$ y $\epsilon_g(T,P)$ como función del número de ejemplos P, se conocen como *curvas de aprendizaje*. Las desviaciones de los valores típicos de las cantidades (2.8), (2.9), (2.10) y (2.11) de sus promedios, se anulan en el límite termodinámico $N \to \infty$. Para que la energía sea una cantidad extensiva, es decir proporcional a N, el número de ejemplos debe estar escaleado con el número total de pesos $P = \alpha N$, donde α debe permanecer finito cuando N crece. Este escaleo garantiza que tanto la energía, como la entropía sean proporcionales a N [8].

La técnica más comúnmente usada para calcular el promedio sobre los ejemplos, es el método de réplicas [14, 15, 16]. Este ha sido muy importante tanto en vidrios de spin, como en redes neuronales y en aquellos casos donde es fácil calcular el promedio de Z, pero no de $\ln Z$, para ello explotamos la identidad

$$\ll \ln Z \gg = \lim_{n\to 0} n^{-1} \ln \ll Z^n \gg, \tag{2.13}$$

donde Z^n es equivalente a una función de partición de n sistemas idénticos no interactuantes caracterizadas por el índice $\gamma = 1, ..., n$. Sin embargo, al calcular el promedio *quenched* sobre los ejemplos, las diferentes réplicas se acoplan. Esto es fácil de ver, construyendo un Hamiltoniano efectivo para las réplicas, intercambiando el orden de las integrales en (2.8), podemos escribir la energía libre como

$$F = -\beta^{-1} \lim_{n\to 0} n^{-1} \ln \int \prod_{\gamma=1}^{n} d\mu(\mathbf{W}_\gamma) \exp\left(-N\alpha H[\mathbf{W}_\gamma]\right), \tag{2.14}$$

donde

$$H[\mathbf{W}_\gamma] = -\ln \int D(\mathbf{S})\, d\mathbf{S} \exp\left[-\beta \sum_{\gamma=1}^{n} \epsilon(\mathbf{W}_\gamma, \mathbf{S})\right]. \tag{2.15}$$

La importancia de esta forma es que H es una cantidad intensiva y no depende del número de ejemplo. Un estudio detallado de este problema puede encontrarse en [17]

2.3 Tratamiento perturbativo

Nuestra meta ahora, consiste en introducir un tratamiento perturbativo, capaz de incorporar los efectos del desorden producido por la aleatoriedad de los ejemplos, en forma aproximada. Nos remitiremos aquí solamente al caso de mayor interés físico, el perceptrón con pesos binarios. En este tipo de redes, presenta un interesante diagrama de fase y los pesos sinápticos W_i solo pueden tomar los valores 1 o -1.

Consideremos en primer lugar, una expansión de H en potencias de β, para luego estudiar los efectos de los diferentes términos de la serie

$$H\left[\mathbf{W}_\gamma\right] = \beta H_0 + \frac{1}{2}\beta^2 H_1 + O\left(\beta^3\right) \tag{2.16}$$

con

$$H_0 = \sum_{\gamma=1}^{n} \varepsilon\left(\mathbf{W}_\gamma\right) \tag{2.17}$$

$$H_1 = \sum_{\delta,\gamma=1}^{n} \left[\varepsilon\left(\mathbf{W}_\gamma\right)\varepsilon\left(\mathbf{W}_\delta\right) - \int d\mu\left(\mathbf{S}\right)\epsilon\left(\mathbf{W}_\gamma,\mathbf{S}\right)\epsilon\left(\mathbf{W}_\delta,\mathbf{S}\right)\right], \tag{2.18}$$

donde H_0 representa la parte no aleatoria de la energía de entrenamiento, mientras que H_1 es el acoplamiento entre dos réplicas diferentes proveniente de la aleatoriedad de los ejemplos. Los términos de orden creciente en β, corresponde al acoplamiento entre más réplicas. Las réplicas pueden ser consideradas como *partículas* con N grados de libertad. El primer termino en (2.16), describe el acoplamiento de las partículas con un campo externo, mientras que el segundo representa la interacción entre dos de estas partículas por un potencial efectivo que depende de la distancia de Hamming entre las réplicas. La temperatura T, está asociada con la constante de acoplamiento. Si bien a bajas temperaturas los ordenes crecientes son cada vez más importante, es posible extraer resultados importantes considerando solo los primeros ordenes.

2.3.1 Límite de altas temperatura

En el límite $\beta \to 0$, con $\alpha\beta$ constante, solo H_0 sobrevive, como las fluctuaciones de la energía provenientes de la aleatoriedad de los ejemplos, son del orden $\sqrt{P}$, y por lo tanto pueden ser despreciadas [17]. El límite de alta temperatura es interesante debido a que es capaz de dar, sin mucha labor algebráica ni numérica, una descripción cualitativa sobre las transiciones de fases a estados generalización perfecta con $R = 1$. El proceso de aprendizaje puede ser comprendido como un proceso dinámico con una energía efectiva $P\varepsilon$. El cálculo explícito de la función de generalización (ver Apéndice A) da

$$\varepsilon(R) = \frac{1}{4\pi} \int \frac{dxdy}{\sqrt{1-R^2}} \exp\left[-\frac{x^2+y^2-2xyR}{2(1-R^2)}\right] [g(x) - g(y)]^2 \tag{2.19}$$

donde R es el solape entre el vector peso del PE y el de PM, es decir, $R_\gamma = N^{-1}\mathbf{W}_\gamma \cdot \mathbf{W}_0$. La energía efectiva (2.19) es más suave que la E_t puesto que es el promedio sobre todo el espacio de entradas y no solo sobre el CE. Estudiaremos ahora, un perceptrón de pesos binarios, con función de transferencia $g(x) = \tanh(j\,x)$, donde j es un parámetro de alinealidad, que interpola entre dos situaciones bien conocidas [17].

Usando la aproximación de simetría de réplica $R_\gamma = R$ y tomando el límite $n \to 0$, podemos escribir la energía libre (2.14) como

$$F \equiv -\beta^{-1} \ln \ll Z^n \gg = \int dR \exp\left[N\left(S(R) - \alpha H_0(R)\right)\right], \tag{2.20}$$

donde S es el logaritmo de la densidad de redes con solape R dada por

$$S = N^{-1} \ln \int d\mu(\mathbf{W})\, \delta\left(NR_\gamma - \mathbf{W}\cdot\mathbf{W}_0\right), \tag{2.21}$$

introduciendo la variable auxiliar $\widehat{R}$, reescribimos (2.21) como

$$S = N^{-1} \ln \int_{-i\infty}^{i\infty} \frac{d\widehat{R}}{2\pi i} \exp\left[-NR\widehat{R} + \ln \int d\mu(\mathbf{W})\, e^{\widehat{R}\mathbf{W}\cdot\mathbf{W}_0}\right]. \tag{2.22}$$

Tomando el límite termodinámico $N \to \infty$ y reemplazando $\int d\mu(\mathbf{W})$ por $\sum_{W_i=\pm 1}$, obtenemos

$$s = -R\widehat{R} + \ln 2 \cosh \widehat{R} \tag{2.23}$$

Finalmente, la energía libre por neurona está dada por

$$\beta f = \alpha\beta\varepsilon\left(R\right) + R\widehat{R}\ln 2\cosh\widehat{R}. \tag{2.24}$$

Extremalizando con respecto a las variables R y $\widehat{R}$, y eliminando $\widehat{R}$, obtenemos el estado de equilibrio termodinámico. Como podemos ver, las funciones termodinámicas dependen solamente de una temperatura efectiva T/α y tanto ϵ_t como ϵ_g, son iguales.

Por otro lado, observemos que la entropía puede ser escrita en una más familiar

$$s = -\frac{1+R}{2}\ln\left(\frac{1+R}{2}\right) - \frac{1-R}{2}\ln\left(\frac{1-R}{2}\right), \tag{2.25}$$

correspondiente al modelo de Ising, donde R juega ahora el rol de la magnetización.

Consideremos ahora, un perceptrón de salida booleana. Tomando $j \to \infty$, la $\tanh\left(j\,x\right) \to \text{sign}\left(x\right)$ y la función de generalización (2.19), tiene una expresión analítica $\varepsilon = 2/\pi\cos^{-1}\left(R\right)$ [17, 8], y la ecuación de punto de ensilladura concomitante, que determina el estado de equilibrio termodinámico, esta dada por

$$R = \tanh\left(\frac{2\widetilde{\alpha}}{\pi\left(1-R^2\right)^{1/2}}\right), \tag{2.26}$$

donde $\widetilde{\alpha} = \alpha\beta$. La ecuación (2.26) presenta una solución en $R = 1$ para todo valor de $\widetilde{\alpha}$. Existen además otras soluciónes para $\widetilde{\alpha} < \widetilde{\alpha}_{sp} = 1.04$, por arriba de dicho valor, el estado de equilibrio corresponde a un estado estable con $R = 1$. El sistema presenta una transición de fase de primer orden en $\widetilde{\alpha}_{sp}$, correspondiente a la transición *spinodal*. Por otra parte, entre $0.85 = \widetilde{\alpha}_{th} < \widetilde{\alpha} < \widetilde{\alpha}_{sp}$, los estados con $R < 1$ corresponden a estado de generalización pobre y son metaestables. El sistema presenta una transición de fase de segundo orden en $\widetilde{\alpha}_{th}$, correspondiente a la transición termodinámica.

En el caso de un perceptrón de salida lineal $g\left(x\right) = x$, la función de generalización esta dada por $\varepsilon = 1 - R$ por lo que la ecuación de punto de ensilladura es

$$R = \tanh\left(\widetilde{\alpha}\right). \tag{2.27}$$

En este caso, el sistema no presenta transiciones de fase y las curvas de aprendizaje siguen una ley asintótica. Nuestro interés consiste en saber si existe un valor crítico

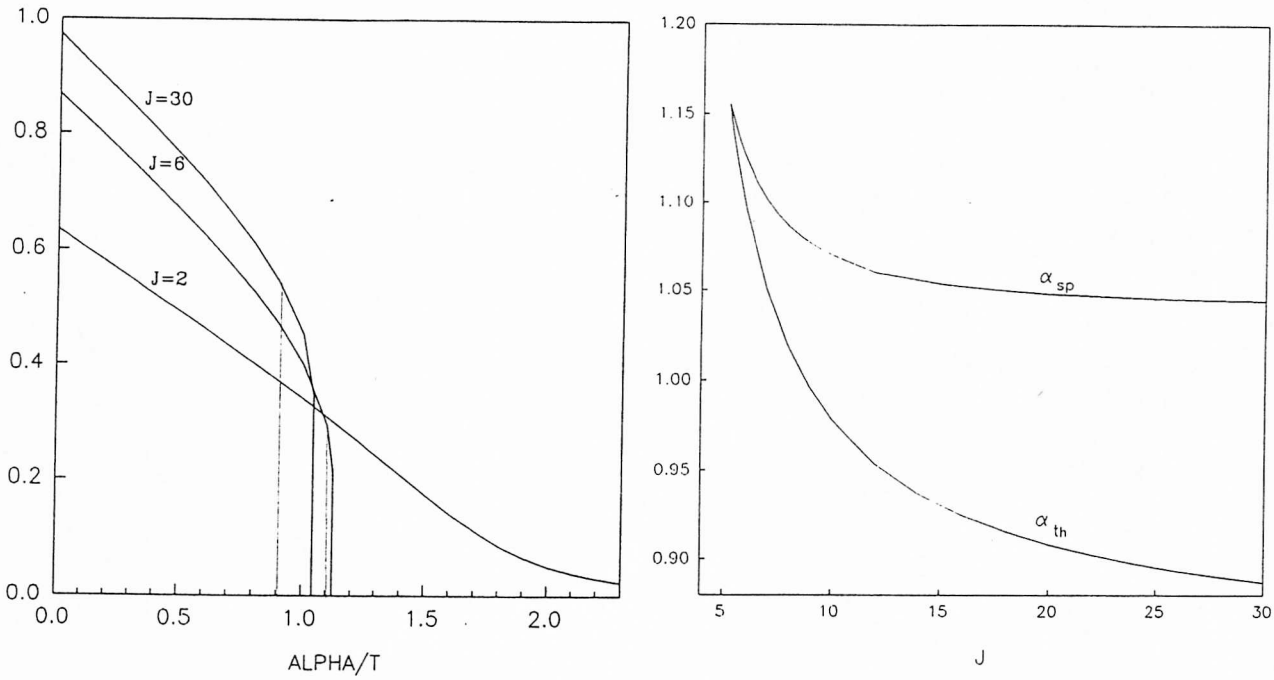

Figura 2.1: a.- El error de generalización versus $\widetilde{\alpha}$. La transición termodinámica está marcada con línea punteada. b.- Comportamiento de $\widetilde{\alpha}_{sp}$ y $\widetilde{\alpha}_{th}$ según j.

del parámetro de alinealidad j para el cual, el sistema realiza o no, una transición de fase al estado de generalización perfecta [18].

En el análisis de un perceptrón con una función de transferencia sigmoidal como $\tanh(j\ x)$ [18] no es posible encontrar expresiones cerradas como en (2.26) y (2.27), y es necesario un estudio numérico del problema. Encontramos que para valores de $j > j_c = 5.2$ el comportamiento termodinámico es similar al perceptrón de salida booleana (Fig. 2.1). Es decir que mientras $\widetilde{\alpha} < \widetilde{\alpha}_{sp}$, la energía libre tiene dos mínimos, uno en $R = 1$ (mínimo local) y otro entre $0 < R < 1$ (mínimo global). Para $\widetilde{\alpha} > \widetilde{\alpha}_{th}$, los estados con $0 < R < 1$, corresponden a un mínimo local con generalización pobre. Ambos mínimos están separados por una barrera escaleada con N. Así la red estudiante evoluciona rápidamente hasta el mínimo entre $0 < R < 1$ y luego lentamente alcanza el mínimo global en $R = 1$, es decir el estado de generalización perfecta (ver Fig. 2.2). Sin embargo, para $\widetilde{\alpha} > \widetilde{\alpha}_{sp}$ el mínimo entre $0 < R < 1$ desaparece y el PE converge rápidamente al estado de generalización perfecta $R = 1$.

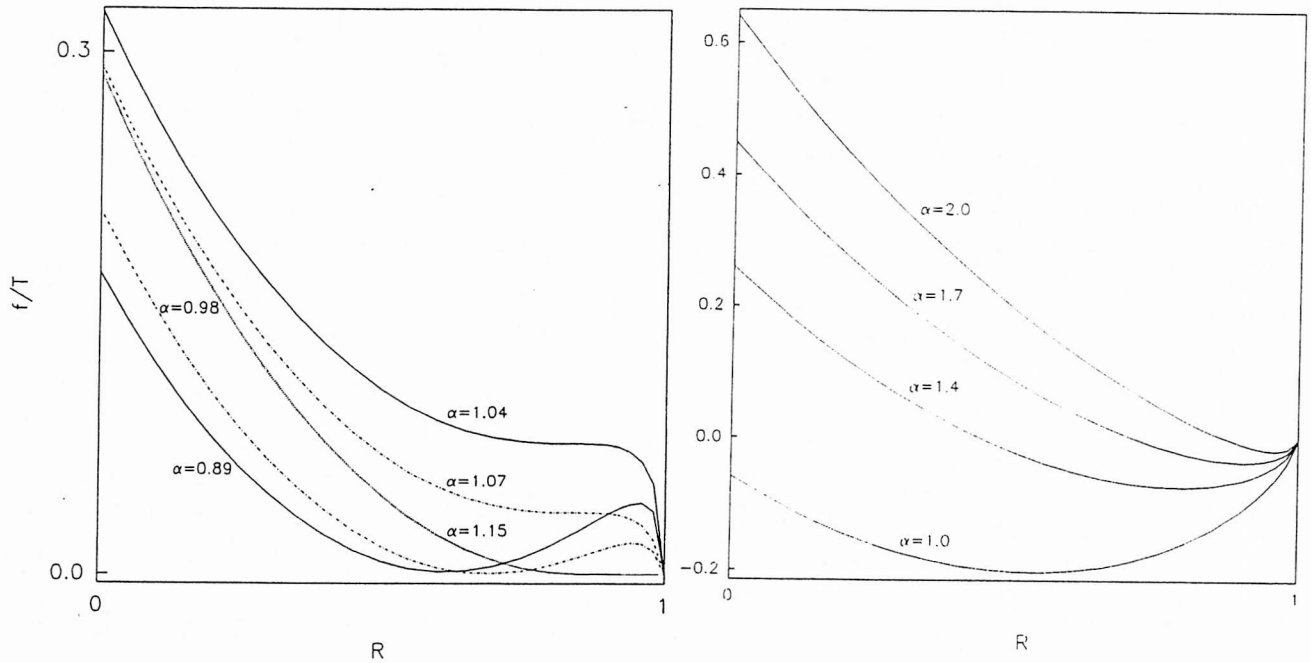

Figura 2.2: a.- Energía libre βf como función de R para $j \geq j_c$. Para $\widetilde{\alpha} = \widetilde{\alpha}_{sp}$ y $\widetilde{\alpha} = \widetilde{\alpha}_{th}$. La línea sólida corresponde a $j = 30$, la línea cortada corresponde a $j = 10$ y la línea punteada es con $j = j_c = 5.2$. En este régimen existen transiciones de fase. b.- Para $j < j_c$ no hay transiciones de fase a ningún valor de $\widetilde{\alpha}$. $\widetilde{\alpha} = 1.2$ y 1.4.

A medida que el parámetro j se acerca al valor crítico j_c, $\widetilde{\alpha}_{sp}$ disminuye, mientras que $\widetilde{\alpha}_{th}$ crece hasta encontrase en $\widetilde{\alpha} = 1.15$ cuando j toma el valor 5.2. Para $j < j_c$, el comportamiento termodinámico encontrado es similar a un perceptrón con salida lineal, en el sentido que no existen transiciones de fase sino que el PE se acerca al PM en forma asintótica. En la Fig. 2.3 se puede ver las curvas de aprendizaje para distintos valores j.

Si bien esta aproximación predice la existencia o no, de una transición de fase a un estado de generalización perfecta, no es suficiente para determinar el comportamiento termodinámico a cualquier temperatura, nada puede decir acerca de la existencia de una fase de vidrio de spin a bajas temperatura, como predice la teoría *quenched* completa [17]. En la próxima sección analizaremos la posibilidad de incorporar el desorden producido por la aleatoriedad de los ejemplos, en el marco de la teoría perturbativa.

2.3.2 Perceptrón booleano a segundo orden

Cuando la temperatura desciende el acoplamiento entre las réplicas se vuelve más importante. La razón de este comportamiento la podemos encontrar en que a α finito las fluctuaciones en la energía provenientes de la aleatoriedad de los ejemplos ya no son despreciables. Una forma de considerar el desorden producido por los ejemplos, consisten en introducir la interacción entre dos réplicas[2] [19], incorporando a nuestras consideraciones el siguiente término en el desarrollo (2.16), y calcular la energía de correlación H_1. Para calcular está contribución, tenemos que evaluar la correlación

$$C1_{\gamma\delta} = \int d\mu\left(\mathbf{S}\right)\epsilon\left(\mathbf{W}_\gamma,\mathbf{S}\right)\epsilon\left(\mathbf{W}_\delta,\mathbf{S}\right) \tag{2.28}$$

En el Apéndice A podemos ver el cálculo explícito

$$C1_{\gamma\delta} = \int \frac{dxdydz}{\left(2\pi\right)^{3/2}\,\Xi_{\gamma\delta}} \exp\left[-\frac{K\left(x,y,z\right)}{2\Xi_{\gamma\delta}}\right]\frac{1}{4}\left[g\left(x\right)-g\left(z\right)\right]^2\left[g\left(y\right)-g\left(z\right)\right]^2, \tag{2.29}$$

donde

$$\begin{aligned} K\left(x,y,z\right) &= 2xy\left(R_\gamma R_\delta - Q_{\gamma\delta}\right) - 2zx\left(R_\gamma - R_\delta Q_{\gamma\delta}\right) - 2zy\left(R_\delta - R_\gamma Q_{\gamma\delta}\right) \\ &\quad z^2\left(1-Q_{\gamma\delta}^2\right) + y^2\left(1-R_\gamma^2\right) + x^2\left(1-R_\delta^2\right), \qquad (2.30) \\ \Xi_{\gamma\delta} &= \left(1-R_\gamma^2\right)\left(1-R_\delta^2\right) - \left(R_\gamma R_\delta - Q_{\gamma\delta}\right)^2. \qquad (2.31) \end{aligned}$$

En esta aproximación, aparece un nuevo parámetro de orden $Q_{\gamma\delta}$, el cual es una matriz de $n \times n$, cuyos elementos dan el solape entre dos réplicas del sistema, es decir $Q_{\gamma\delta} = N^{-1}\mathbf{W}_\gamma \cdot \mathbf{W}_\delta$. Este parámetro no aparecía en el límite de altas temperaturas, pues no incluimos los acoplamientos entre réplicas. Pero a medida que la temperatura desciende estos acoplamientos pueden conducir a la aparición de fases cualitativamente diferentes, como la de vidrio de spin. Para describir correctamente esta fase, la matriz $Q_{\gamma\delta}$ debe tener una estructura adecuada que discutiremos brevemeneteen la próxima sección. Las expresión (2.29) es completamente general. Pero nosotros estamos interesados en el perceptrón con salida booleana obtenemos. Nuestro próximo paso consiste en calcular la energía libre (2.14) con un Hamiltoniano aproximado

[2]Si bien a bajas temperaturas los ordenes crecientes son cada vez más importantes, podemos esperar resultados aceptables para $T \gtrsim 1$ tomando en cuenta las correlaciones de segundo orden.

que tiene en cuenta la acoplamiento entre dos réplicas, en el perceptrón de salida booleana. En este caso, H' está dado por (ver Apéndice B)

$$H' = \frac{\beta}{\pi}\sum_{\gamma}^{n}\cos^{-1}(R_\gamma) + \frac{\beta^2}{4\pi}\sum_{\gamma\neq\delta}^{n}\left[\pi/2 - \tan^{-1}\left(\frac{Q_{\gamma\delta}}{(1-Q_{\gamma\delta}^2)^{1/2}}\right)\right], \tag{2.32}$$

donde hemos eliminados el primer terminos de (2.18) por ser de orden n^2. El Hamiltoniano (2.32), depende de los pesos solamente através de los parámetros de orden R_γ y $Q_{\gamma\delta}$. Por esta razón para calcular la energía libre tenemos que introducir las variables de integración $\widehat{R}_\gamma$ y $\widehat{Q}_{\gamma\delta}$

$$\begin{aligned}
\ll Z^n \gg &= \int\prod_{\gamma<\delta} dQ_{\gamma\delta}\int\prod_{\gamma} dR_\gamma \exp\left(-N\alpha H'\left[R_\gamma, Q_{\gamma\delta}\right]\right)\times \\
&\quad \int\prod_{\gamma} d\mu(\mathbf{W}_\gamma)\prod_{\gamma<\delta}\delta\left(NQ_{\gamma\delta} - \mathbf{W}_\gamma\cdot\mathbf{W}_\delta\right)\prod_{\gamma}\delta\left(NR_\gamma - \mathbf{W}_\gamma\cdot\mathbf{W}_0\right) \\
&= \int\prod_{\gamma<\delta}\frac{dQ_{\gamma\delta}d\widehat{Q}_{\gamma\delta}}{2\pi i}\int\prod_{\gamma}\frac{dR_\gamma d\widehat{R}_\gamma}{2\pi i}\exp\left(-N(\alpha H' - S)\right),
\end{aligned} \tag{2.33}$$

donde S es el logaritmo de la densidad de redes con solapes R_γ y $Q_{\gamma\delta}$

$$\begin{aligned}
S &= N^{-1}\ln\int\prod_{\gamma} d\mu(\mathbf{W}_\gamma)\exp\left(\sum_{\gamma}\widehat{R}_\gamma\mathbf{W}_\gamma\cdot\mathbf{W}_0 + \sum_{\gamma<\delta}\widehat{Q}_{\gamma\delta}W_\gamma\cdot\mathbf{W}_\delta\right) \\
&\quad -\sum_{\gamma}R_\gamma\widehat{R} - \sum_{\gamma<\delta}Q_{\gamma\delta}\widehat{Q}_{\gamma\delta}.
\end{aligned} \tag{2.34}$$

En el límite termodinámico $N\to\infty$, la integral (2.33) está dominada por el mínimo en las variables R_γ,$\widehat{R}_\gamma$, $Q_{\gamma\delta}$ y $\widehat{Q}_{\gamma\delta}$, por lo tanto la energía libre se puede escribir como

$$\begin{aligned}
-\beta f &= \lim_{n\to 0}\frac{1}{n}N^{-1}\ln\ll Z^n\gg \\
&= \min\{\alpha H'\left[R_\gamma, Q_{\gamma\delta}\right] - S\left[R_\gamma, \widehat{R}_\gamma, Q_{\gamma\delta}, \widehat{Q}_{\gamma\delta}\right]\}.
\end{aligned} \tag{2.35}$$

Para evaluar la ecuación (2.35) es necesario hacer alguna consideración acerca de los parámetros de orden. La propuesta más simple es la de simetría de réplica que considera que todas las réplicas tienen el mismo solape con el PM y que el solape entre las diferentes réplicas es el mismo para todo par de ellas

$$\begin{aligned}
Q_{\gamma\delta} &= \delta_{\gamma\delta} + (1-\delta_{\gamma\delta})\,q, \\
R_\gamma &= R.
\end{aligned} \tag{2.36}$$

La interpretación física del parámetro de orden q es similar al parámetro de orden de Edwards-Anderson en vidrios de spin [14, 20, 21], caracteriza el solape típico entre dos mínimos de la energía E_t. Cuando α aumenta las diferentes soluciones se vuelven más correlacionadas por lo tanto q se acerca a 1. Para $\alpha = \alpha_c$, resulta $q = 1$. En este punto es necesario aclarar que la propuesta de simetría de réplica, constituye una aproximación válida si las temperaturas no son muy bajas.

Tomando el límite $n \to 0$ y usando la ecuacion (2.12) podemos escribir

$$f = \alpha\epsilon_t - Ts. \tag{2.37}$$

donde

$$\epsilon_t = \frac{1}{\pi}\cos^{-1}(R) - \frac{\beta}{4\pi}\left[\frac{\pi}{2} - \tan^{-1}\left(\frac{q}{(1-q^2)^{1/2}}\right)\right], \tag{2.38}$$

$$s = \int Dz \ln \int d\mu(\mathbf{W}) \exp\left[\mathbf{W}\cdot\left(\sqrt{\widehat{q}}\mathbf{z} + \mathbf{W}_0\widehat{R}\right)\right] - \frac{1}{2}(1-q)\,\widehat{q} - R\widehat{R} + . \tag{2.39}$$

Las variables $R, \widehat{R}, q$ y $\widehat{q}$, deben ser determinadas en forma autoconsistente a partir de las ecuaciones de punto de ensilladura de la energía libre. Una caracteristica a destacar es que a diferencia de altas temperaturas, el error de entrenamiento ϵ_t y el error de generalización ϵ_g son distintos. En esta aproximación los estos difieren en una cantidad proporcional a β [ver ecuación (2.38)]. Para poder continuar es necesario especificar, el tipo de restricción que opera sobre los pesos sinápticos. Consideremos el de pesos binarios. En este caso, la medida en el espacio de los pesos esta dada por

$$d\mu(\mathbf{W}) = \prod_i dW_i\left(\delta(W_i - 1) + \delta(W_i + 1)\right). \tag{2.40}$$

En este caso el cálculo de la entropía de (2.39) da

$$s = -\frac{1}{2}(1-q)\,\widehat{q} - R\widehat{R} + \int Dz \ln 2\cosh\left[\left(\sqrt{\widehat{q}}z + \widehat{R}\right)\right]. \tag{2.41}$$

Extremalizando f con respecto a los parámetros $R, \widehat{R}, q$ y $\widehat{q}$ y eliminando $\widehat{R}$ y $\widehat{q}$ de las ecuaciones, obtenemos

$$R = \int Dz \tanh\left(\sqrt{\frac{\alpha\beta^2}{2\pi}\frac{1}{(1-q^2)^{1/2}}}z + \frac{\alpha\beta}{\pi}\frac{1}{\sqrt{1-R^2}}\right),$$

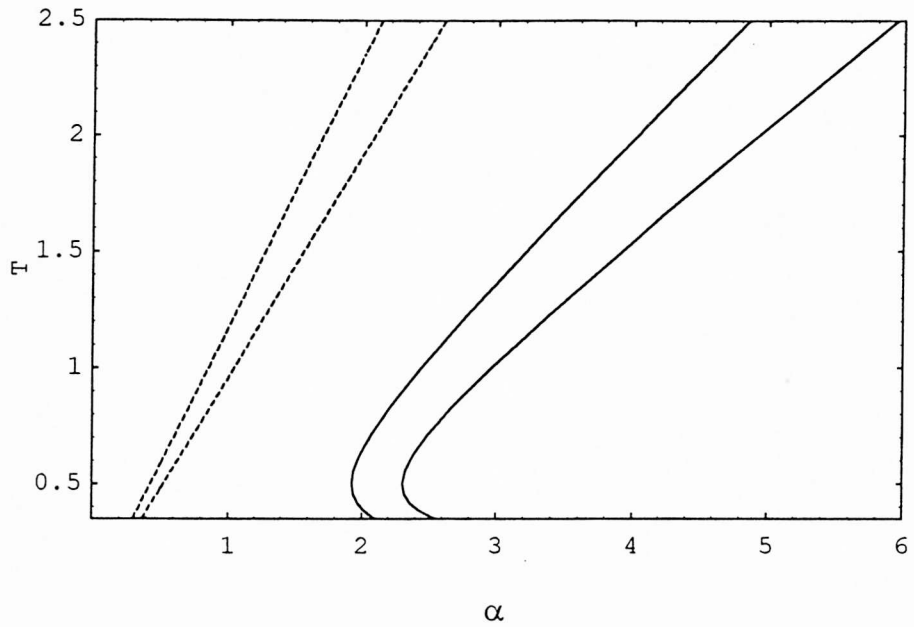

Figura 2.3: Diagrama de fase obtenido a primer (línea de trazos) y a segundo orden (línea continua.)

$$q = \int Dz \tanh^2 \left(\sqrt{\frac{\alpha\beta^2}{2\pi} \frac{1}{(1-q^2)^{1/2}}} z + \frac{\alpha\beta}{\pi} \frac{1}{\sqrt{1-R^2}} \right). \tag{2.42}$$

Las ecuaciones (2.42), junto con la exigencia de que el Hessiano $[f(R,q)] = 0$, describen la línea espinodal en diagrama fase (T, α) (Fig. 2.3). En el límite $\beta \to 0$ con $\alpha\beta$ constante recuperamos los resultados de alta temperatura, con una nueva e interesante relación $q = R^2$ [19], (la que podría interpretarse como un campo medio). Esta relación no podría aparecer en el tratamiento a primer orden, pues no dice nada concerniente a q.

La Fig. 2.4 muestra las curvas de aprendizaje a $T = 1$, la transición spinodal ocurre a $\alpha_c = 2.95$, en buen acuerdo con los resultados de la teoría *quenched* completa [17], mejorando visiblemente las prediciones de primer orden, en la cual $\alpha_c = 2.08$. La curva de transición termodinámica, determinada por las ecuaciones (2.42) y $f(R,q) = 0$, separa estos estados de aquellos de generalización pobre que corresponde a un mínimo global de la energía libres. La transición termodinámica está marcada con línea de punto en $\alpha = 2.438$.

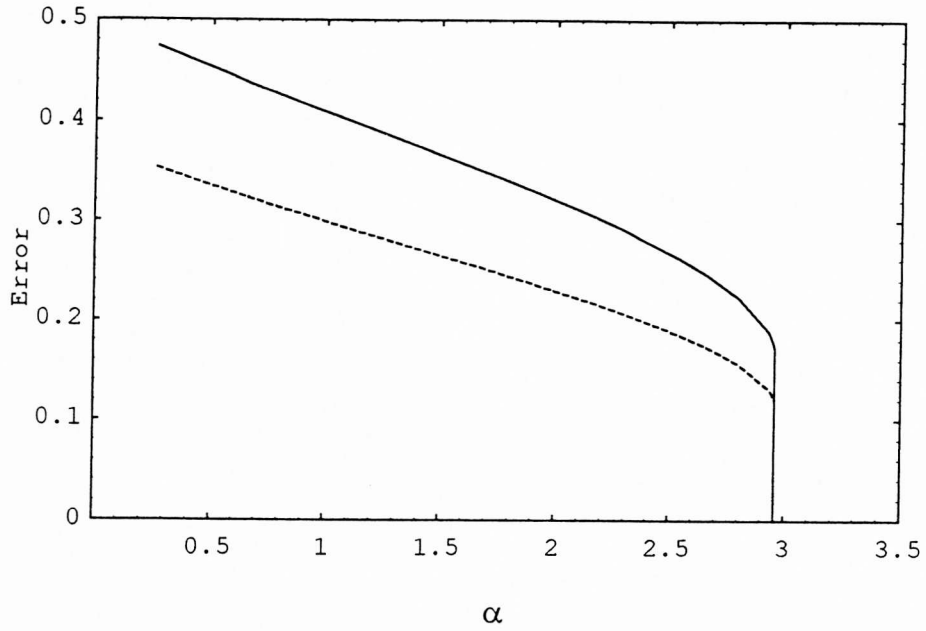

Figura 2.4: Error de generalización (línea llena) y error de entrenamiento (línea de trazo), obtenidas a segundo orden.

2.3.3 La fase de vidrios de spin

Puesto que el Hamiltoniano (2.15) es invariante bajo permutación de los índices de réplicas, podríamos esperar que la aproximación de simetría de réplica dada por (2.36) describa el sistema adecuadamente. Sin embargo, en ciertas condiciones[3], puede ocurrir una rotura espontánea de la simetría de réplica indicando una fase de vidrio de spin. Para un sistema con un número discreto de grado de libertad la entropía del sistema debe ser no negativa, por lo que establece un criterio de validez de la propuesta de simetría de réplica. Este criterio de entropía cero provee una cota inferior para la temperatura a la cual ocurre la rotura de simetría. Para estudiar el sistema a bajas temperaturas, donde la simetría de réplica ya no es válida, es necesario dar al parámetro de orden $Q_{\gamma\delta}$ una estructura más compleja, en el marco del tratamiento completo de la teoría de réplica [17], que incorpore todos los ordenes en la expansión (2.16).

La fase de vidrios de spin está caracterizada por que los valores de espectación

[3]Estrictamente cuando la temperaturas y el número de ejemplos por peso es pequeño.

de las correlaciones entre diferentes réplicas tienen una complicada dependencia de los índices de réplicas [22]. Físicamente, la fase de vidrio esta caracterizada por la existencia de un estado fundamental altamente degenerado como resultado de una fuerte frustración en el sistema. Estos estados ocupan regiones desconectadas del espacio de configuración que están separadas por barreras que divergen con N; razón por la cual la evolución dinámica de los W_i se vuelve desconectada y la ergodicidad es rota, conduciendo un aprendizaje anormalmente lento. Como las réplicas son sistemas independientes sujetos al desorden de los ejemplos, diferentes réplicas pueden quedar atrapadas en diferentes regiones del espacio de los pesos, llamados *estados puros* α de peso relativo $P_a = \exp(-\beta F_a)$ [23]. El solape promedio de dos réplicas en el mismo estado puro es diferente del de dos réplicas en dos diferentes estados puros. Siguiendo el esquema de Parisi, trataremos la rotura de simetría en la aproximación *one-step*. La matriz $Q_{\gamma\delta}$ adquiere la estructura de bloques de submatrices $m \times m$, que se obtiene por dividir las n réplicas en n/m grupos de m réplicas. Los elementos de matriz $Q_{\gamma\delta}$ toman los valores q_1 si γ y δ pertenecen al mismo estado puro y q_0 en cualquier otro caso. El significado físico de los parámetros q_1, q_0 y m está dado por

$$\begin{aligned} q_1 &= N^{-1} \ll \langle \mathbf{W}_a \rangle_T^2 \gg \\ q_0 &= N^{-1} \ll \langle \mathbf{W}_a \rangle_T \cdot \langle \mathbf{W}_b \rangle_T \gg \qquad (a \neq b) \\ m &= 1 - \sum_a P_a^2 \end{aligned} \tag{2.43}$$

así m es la probabilidad de encontrar dos copias del sistema en dos estados diferentes. En el límite $n \to 0$, el parámetro m está restringido al rango $0 \leq m \leq 1$. Ahora la energía libre (2.37) para el perceptrón booleano esta dada por

$$\begin{aligned} \epsilon_t &= \frac{1}{\pi}\cos^{-1}(R) - \frac{\beta m}{4\pi}\left[\frac{\pi}{2} - \tan^{-1}\left(\frac{q_0}{\sqrt{1-q_0^2}}\right)\right] \\ &- \frac{\beta(1-m)}{4\pi}\left[\frac{\pi}{2} - \tan^{-1}\left(\frac{q_1}{\sqrt{1-q_1^2}}\right)\right] \\ s &= \frac{1}{2}\left[m\widehat{q}_0 q_0 + (1-m)\,\widehat{q}_1 q_1 - \widehat{q}_1\right] - R\widehat{R} \\ &+ \frac{1}{m}\int Dz_0 \ln \int Dz_1 \left[2\cosh\left(z_0\sqrt{\widehat{q}_0} + z_1\sqrt{\widehat{q}_1 - \widehat{q}_0} + \widehat{R}\right)\right]^m . \end{aligned} \tag{2.44}$$

Buscamos ahora una solución de la ecuación de punto de ensilladura con la propiedad $q_1 = 1$ y $\widehat{q}_1 = \infty$ [17]. Para ello tomamos el límite $q_1 \to 1$, $\widehat{q}_1 \to \infty$ en las ecuaciones (2.44) manteniendo β finito. Obtenemos

$$\begin{aligned} \epsilon_t &= \frac{1}{\pi}\cos^{-1}(R) - \frac{\beta m}{4\pi}\left[\frac{\pi}{2} - \tan^{-1}\left(\frac{q_0}{\sqrt{1-q_0^2}}\right)\right], \\ s &= \frac{1}{m}\left[m^2\widehat{q}_0(q_0-1)\,q - m\widehat{R} + \frac{1}{m}\int Dz \ln\left[2\cosh\left(mz\sqrt{\widehat{q}_0} + m\widehat{R}\right)\right]\right] \end{aligned} \quad (2.45)$$

comparando estas ecuaciones con (2.38) y (2.39), observamos

$$f_{RSB}\left(q_0, \widehat{q}_0, R, \widehat{R}, m, \beta\right) = \frac{1}{m} f_{RS}\left(q_0, m^2\widehat{q}_0, R, m\widehat{R}, m\beta\right) \quad (2.46)$$

en forma similar a los resultados obtenidos en [17] y en [23] para el problema de memorizar patrones al azar en un perceptrón. Un análisis cuantitativo más profundo de la fase de vidrios de spin, deberia ser estudiada con la teoría *quenched* completa, puesto que a $T = O$, todos los términos de las expansión deben ser incorporados. Sin embargo, aún en esta aproximación es posible apreciar cierta estructura caracteristica de la rotura de simetría [19].

2.4 Tratamiento perturbativo para el aprendizaje con ejemplos incorrectos

Para redes tipo perceptrón, el método de réplicas ha provisto un marco para el estudio del aprendizaje de una regla, a partir de un conjunto de ejemplos [17, 18]. En la presente sección, deseamos estudiar los efectos sobre las propiedades termodinámicas de la red cuando el ruido está presente en el CE [19]. En este caso, el ruido se debe al hecho de que solo una parte de los ejemplos ("buenos" ejemplos), son producidos por un PM con la misma arquitectura que el PE, mientras que el resto son seleccionados al azar ("malos" ejemplos). En tal situación, el error de entrenamiento nunca puede anularse. La pregunta a ser respondida aquí es: puede un PE, entrenado bajo estas condiciones "responder" correctamente las "preguntas" del PM? En otras palabras, es aprendible la regla subyacente a los ejemplos provenientes del PM ?

Estudiaremos este problema con el esquema de la Mecánica Estadística, anteriormente expuesto, dentro del marco de una aproximación perturbativa a segundo orden [19], en el caso donde la salida del perceptrón es booleana. Los P ejemplos están dados por

$$\{\mathbf{S}^l, t^l\} \qquad\qquad l = 1, ..., P. \tag{2.47}$$

y asumimos que un subconjunto de P_1 ejemplos son generados por el PM, por lo tanto $t^l = \sigma_0(\mathbf{S}^l)$, con $l = 1, ..., P_1$, mientras los restantes $P_2 = P - P_1$ salidas t^l, son de naturaleza aleatoria. Tanto las entradas $\mathbf{S}^l$ como las salidas t^l son elegidas al azar con una probabilidad $D(\mathbf{S})$ y $D(t)$, desde el espacio de las entradas y las salidas, respectivamente.

La energía E_t, esta definida ahora por

$$E_t(\mathbf{W}) = \sum_{l=1}^{P} \epsilon\left(\mathbf{W}, \mathbf{S}^l, t^l\right), \tag{2.48}$$

donde $\epsilon(\mathbf{W}, \mathbf{S}, t) = \theta\left(-N^{-1/2}(\mathbf{W}\cdot\mathbf{S})\ t\right)$, siendo θ la función de Heaviside.

Ahora el promedio *quenched*, debe ser calculado sobre todo el espacio entrada–salida[4], es decir $\ll ... \gg \equiv \int ... \prod_l d\mu(\mathbf{S}^l) d\mu(t^l)$. La energía libre está dada por

$$F = -\beta^{-1} \lim_{n\to 0} n^{-1} \ln \int \prod_{\gamma=1}^{n} d\mathbf{W}_\gamma \exp\left[-N\alpha\left(\frac{1}{1+\rho} G[\mathbf{W}_\gamma] + \frac{\rho}{1+\rho} G'[\mathbf{W}_\gamma]\right)\right], \tag{2.49}$$

donde $\rho = \frac{P_2}{P_1}$. En (2.49), G establece la contribución de los "buenos ejemplos" y esta dada por

$$G[\mathbf{W}_\gamma] = -\ln \int D(\mathbf{S})\, d\mathbf{S} \exp\left[-\beta \sum_{\gamma=1}^{n} \epsilon(\mathbf{W}_\gamma, \mathbf{S})\right]. \tag{2.50}$$

Como las salidas están asociadas con las entradas a traves del PM, tenemos que $\epsilon(\mathbf{W}_\gamma, \mathbf{S}) = \theta\left(-N^{-1/2}(\mathbf{W}_\gamma\cdot\mathbf{S})(\mathbf{W}_0\cdot\mathbf{S})\right)$ y no es necesario promediar sobre t.

Del mismo modo, la contribución de los "malos ejemplos" está dada por

$$G'[\mathbf{W}_\gamma] = -\ln \int\int D(t)\, dt D(\mathbf{S})\, d\mathbf{S} \exp\left[-\beta \sum_{\gamma=1}^{n} \epsilon(\mathbf{W}_\gamma, \mathbf{S}, t)\right]. \tag{2.51}$$

[4]El error de generalización (2.11) debe ser calculado solo sobre las entradas, puesto que la salidas estan asociadas a estas entradas por ser ejemplos miembros de la regla.

Expandiendo $H = \frac{1}{1+\rho} G[\mathbf{W}_\gamma] + \frac{\rho}{1+\rho} G'[\mathbf{W}_\gamma]$ en potencias de β, y quedandonos a segundo orden, obtenemos

$$H'[\mathbf{W}_\gamma] = \beta H_0 + \frac{1}{2}\beta^2 H_1. \tag{2.52}$$

H_0 representa la parte no aleatoria de la energía, y está dado por

$$H_0 = \frac{1}{1+\rho} \sum_{\gamma=1}^{n} \varepsilon(\mathbf{W}_\gamma) + \frac{\rho}{1+\rho} \sum_{\gamma=1}^{n} \int D(t)\, dt D(\mathbf{S})\, d\mathbf{S} \epsilon(\mathbf{W}_\gamma, \mathbf{S}, t). \tag{2.53}$$

La función de generalización $\varepsilon(R) = \frac{1}{\pi}\cos^{-1}(R)$ como vimos anteriormente, solo depende del parámetro de orden R. Por otra parte, el segundo término del lado derecho de (2.53) es igual a la unidad, (Apéndice C). La contribución de H_1 viene dada por

$$\begin{aligned} H_1 &= \frac{1}{1+\rho} \sum_{\gamma,\delta=1}^{n} \left[\varepsilon(\mathbf{W}_\gamma)\varepsilon(\mathbf{W}_\delta) - \int d\mathbf{S} D(\mathbf{S}) \epsilon(\mathbf{W}_\gamma, \mathbf{S}) \epsilon(\mathbf{W}_\delta, \mathbf{S}) \right] \\ &+ \frac{\rho}{1+\rho} \sum_{\gamma,\delta=1}^{n} \left[1 - \int D(t)\, dt D(\mathbf{S})\, d\mathbf{S} \epsilon(\mathbf{W}_\gamma, \mathbf{S}, t) \epsilon(\mathbf{W}_\delta, \mathbf{S}, t) \right]. \end{aligned} \tag{2.54}$$

A partir de (2.54), podemos ver que las integrales a considerar son

$$C1_{\gamma\delta} = \int d\mathbf{S} D(\mathbf{S}) \epsilon(\mathbf{W}_\gamma, \mathbf{S}) \epsilon(\mathbf{W}_\delta, \mathbf{S}), \tag{2.55}$$

$$C2_{\gamma\delta} = \int dt D(t) D(\mathbf{S})\, d\mathbf{S} \epsilon(\mathbf{W}_\gamma, \mathbf{S}, t) \epsilon(\mathbf{W}_\delta, \mathbf{S}, t). \tag{2.56}$$

La primera integral ya fue analizada en el Apendice B, la integral siguiente considera la energía de correlación que producen los ejemplos "malos". En el Apéndice C, mostramos que estas contribuciones son similares a $C1_{\gamma\delta}$. Eliminando los términos de orden n^2, el Hamiltoniano (2.52) se puede escribir como

$$H' = \frac{\beta}{\pi(1+\rho)} \sum_{\gamma=1}^{n} \cos^{-1}(R_\gamma) + \frac{n\beta\rho}{1+\rho} - \frac{\beta^2}{2(1+\rho)} \sum_{\gamma,\delta=1}^{n} C1_{\gamma\delta} - \frac{\rho\beta^2}{2(1+\rho)} \sum_{\gamma,\delta=1}^{n} C2_{\gamma\delta}. \tag{2.57}$$

Ahora el Hamiltoniano replicado (2.57), depende de los pesos sólo a través de los parámetros de orden R_γ y $Q_{\gamma\delta}$. Introduciendo las variables auxiliares $\widehat{R}_\gamma$ y $\widehat{Q}_{\gamma\delta}$

obtenemos

$$\ll Z^n \gg = \int \prod_{\gamma<\delta} \frac{dQ_{\gamma\delta} d\widehat{Q}_{\gamma\delta}}{2\pi i} \int \prod_{\gamma} \frac{dR_\gamma d\widehat{R}_\gamma}{2\pi i} \exp\left(-N(\alpha H' - S)\right), \tag{2.58}$$

En el límite termodinámico $N \to \infty$, la integral (2.58) sobre las variables $R_\gamma, \widehat{R}_\gamma$, $Q_{\gamma\delta}$ y $\widehat{Q}_{\gamma\delta}$, están dominadas por el punto de ensilladura de la energía libre en las variables R_γ y $Q_{\gamma\delta}$. La energía libre es obtenida por continuación analítica a $n = 0$

$$\begin{aligned} -\beta f &= \lim_{n\to 0} \frac{1}{n} N^{-1} \ln \ll Z^n \gg \\ &= \text{ext}\{\alpha H\left[R_\gamma, Q_{\gamma\delta}\right] - S\left[R_\gamma, \widehat{R}_\gamma, Q_{\gamma\delta}, \widehat{Q}_{\gamma\delta}\right]\}. \end{aligned} \tag{2.59}$$

Tomando el límite $n \to 0$ en (2.57) y reemplazando $C1_{\gamma\delta}$ y $C2_{\gamma\delta}$, obtenemos una expresión para el error de entrenamiento. Usando la ecuación (2.12) podemos escribir

$$f = \alpha\epsilon_t - Ts, \tag{2.60}$$

donde

$$\epsilon_t = \frac{1}{\pi(1+\rho)} \cos^{-1}(R) + \frac{\rho}{1+\rho} - \frac{\beta}{4\pi}\left[\frac{\pi}{2} - \tan^{-1}\left(\frac{q}{(1-q^2)^{1/2}}\right)\right] \tag{2.61}$$

$$s = -\frac{1}{2}(1-q)\,\widehat{q} - R\widehat{R} + \int Dz \ln 2\cosh\left[\left(\sqrt{\widehat{q}}z + \widehat{R}\right)\right]. \tag{2.62}$$

Podemos observar en (2.61) las diferentes contribuciones al error de entrenamiento, el primer y segundo término constituyen el aporte a primer orden en β, las contribuciones de los buenos ejemplos y de los malos ejemplos están separadas y aparecen pesadas por un factor que involucra la razón P_2/P_1 [19]. El último término en (2.61) corresponde a las contribuciones de segundo orden en β, en este caso tanto el aporte de los buenos como los malos ejemplos tienen la misma forma, esto es razonable puesto que la aleatoriedad está presente en ambos subconjuntos de ejemplos.

Los valores de equilibrio de, $R, \widehat{R}, q$ y $\widehat{q}$ deben ser determinados en forma autoconsistente, a partir de las ecuaciones de punto de ensilladura de la energía libre. Extremalizando (2.60) con respecto a los parámetros $R, \widehat{R}, q$ y $\widehat{q}$ y eliminando $\widehat{R}$ y $\widehat{q}$,

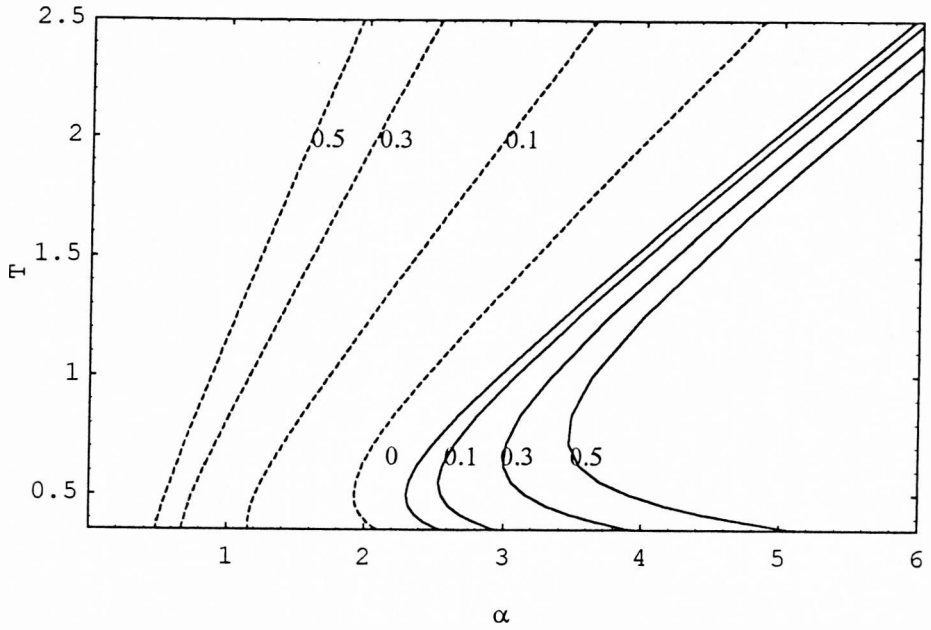

Figura 2.5: Diagrama de fase (α, T) obtenido con la aproximación de segundo orden (en β) para diferentes valores del parámetro de ruido ρ. Las líneas llenas corresponden a las transiciones spinodales y las líneas cortadas a la termodinámica.

obtenemos las ecuaciones pertinentes

$$\begin{aligned} R &= \int Dz \tanh\left[\sqrt{\frac{\beta^2\alpha}{2\pi}\frac{1}{(1-q^2)^{1/2}}}z + \frac{\alpha\beta}{\pi(1+\rho)}\frac{1}{\sqrt{1-R^2}}\right] \\ q &= \int Dz \tanh^2\left[\sqrt{\frac{\beta^2\alpha}{2\pi}\frac{1}{(1-q^2)^{1/2}}}z + \frac{\alpha\beta}{\pi(1+\rho)}\frac{1}{\sqrt{1-R^2}}\right]. \end{aligned} \tag{2.63}$$

Para $T \gtrsim 0.5$, existe una transición de fase de primer orden, desde un estado con generalización pobre a un estado con $R = 1$, cuando α alcanza cierto valor crítico α_{sp}. Este estado corresponde a una fase de generalización perfecta y es alcanzado aún cuando ρ constituye una fracción apreciable de la unidad. En la Fig. 2.5, podemos ver el diagrama de fase para diferentes niveles de ruido $\rho = 0, 0.1, 0.3$ y 0.5. Encontramos que a medida que ρ aumenta α_{sp} es mayor. Existe una región del diagrama de fase donde los estados de generalización pobre son metaestable. La curva de transición termodinámica, determinada por las ecuaciones (2.63) y $f(R, q) = 0$, separa estos estados de aquellos de generalización pobre que corresponde a un mínimo global de la energía libres. La curva de transición cambia apreciablemente con ρ. El valor crítico

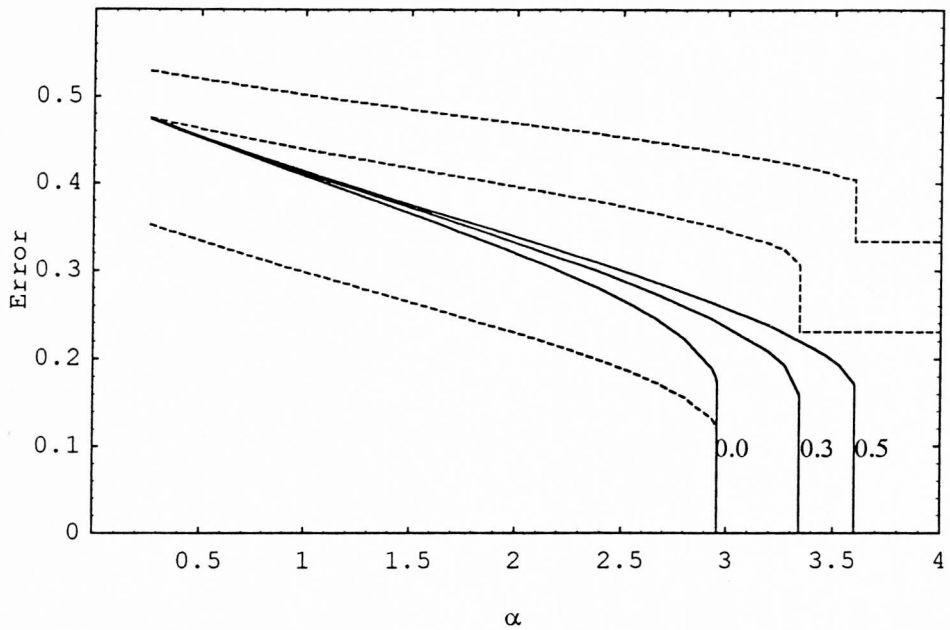

Figura 2.6: Curvas de aprendizaje para diferentes niveles de ρ a $T = 1$ en la aproximación de segundo orden. El error de generalización aparece en línea llena, mientras que el error de entrenamiento corresponde a la línea cortada.

de α para las transiciones termodinámicas, α_{th} es considerablemente más pequeño, por lo que los estados metaestables de generalización pobre son bastantes más pobres. Las anomalías en el diagrama de fase a bajas temperaturas ($T < 0.5$ cuando $\rho = 0$), son un efecto de la aproximación. Estas anomalías son más notorias cuando el ruido aumenta.

El error de entrenamiento está dado por (2.61) y como era de esperar, no se anula cuando el sistema realiza una transición de fase [19]. Esto ocurre porque no existe un vector **W** que sea solución de (2.47). Si consideramos el error de generalización como un promedio restringido sobre los pares preguntas–respuestas del PM, entoces el error pertinente está dado por $\epsilon_g = \frac{1}{\pi}\cos^{-1}(R)$. En la Fig. 2.6. podemos ver las curvas de aprendizaje para diferentes valores ρ, a $T = 1$.

Si hacemos ρ= 0, en (2.61) y (2.63) obtenemos una aproximación de segundo orden al perceptrón de pesos binarios entrenado sin ruido. En la Fig. 2.6, podemos ver que a $T = 1$, la transición espinodal ocurre a $\alpha_{sp} = 2.95$, en perfecto acuerdo a los resultados previos. En la Fig. 2.7, podemos ver el comportamiento de R tanto

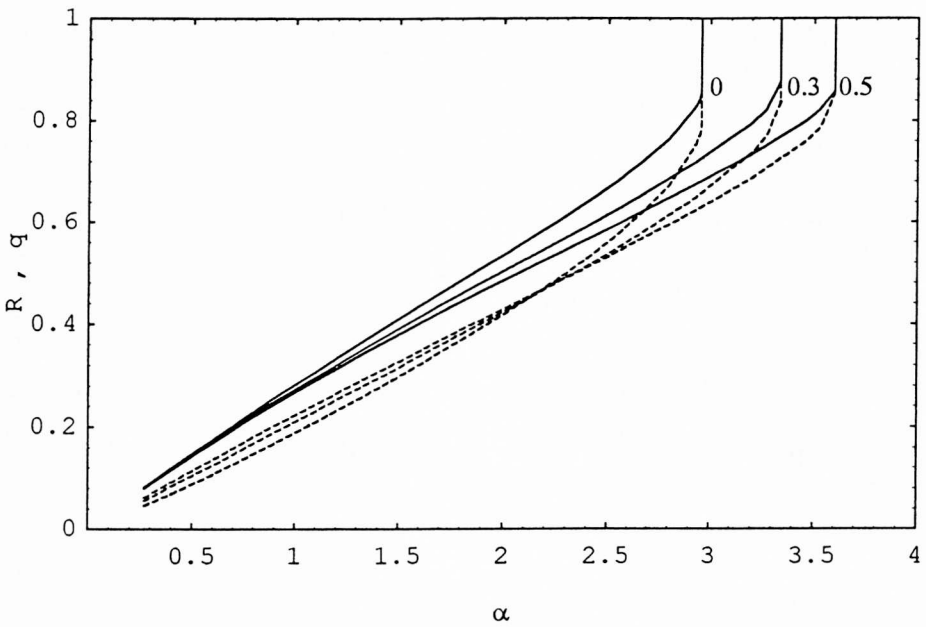

Figura 2.7: Comportamiento de los parámetros de orden R (línea de trazos) y q (línea llena) según α para $T = 1$.

como de q segun varia α. Cuando α aumenta, más y más correlacionados están las diferentes soluciones y q se aproxima a la unidad. Para $\alpha = \alpha_{sp}$, tenemos $q = 1$ y la degeneración es rota.

2.5 Conclusiones

Hemos investigado, dentro de un esquema de aprendizaje visto como un proceso de relajación, diferentes aspectos de la termodinámica del aprendizaje de un perceptrón. En lo referente al perceptrón no lineal, hemos descubierto que existe un valor crítico j_c para el cual dicho perceptrón tiene un comportamiento similar al de salida booleana, es decir presenta transición de fase a altas temperatura.

Por otro lado, hemos presentado un nuevo tratamiento perturbativo que incorpora los acoplamientos entre dos réplicas, introduciendo el parámetro de orden q, solo presente en la teoría *quenched*. Este esquema nos permite estudiar con bastante exactitud diverso problemas, dentro del rango de validez de la simetría de réplica.

Por último, concluimos a partir de nuestra investigación que aún con ejemplos

ruidoso el perceptrón puede alcanzar el estado de generalización perfecta, es decir, que el perceptrón de pesos bianrios es capaz de aprender la regla subyacente a los buenos ejemplos provistos por el PM. Además los estados metastable de generalización pobre son mucho más pobres que en el aprendizaje sin ruido. De esta manera hemos introducido una técnica perturbativa novedosa como herramienta para el estudio de la termodinámica de los procesos de aprendizaje cuando la temperaturas del proceso de aprendizaje están por arriba de 0.5.

Bibliografía

[1] D.E. Rumelhart y J.L. McClelland, *Parallel Distributed Processing* (MIT, Cambridge, MA., 1986).

[2] P. del Giudice, S. Franz, y M.A. Virasoro, J. Phys. (Paris) **50**, 121 (1989).

[3] G. Gyorgyi y N. Tishby, *Neural Networks and Spin Glasses*, editado por W.K. Theumann y R. Koberle (World Scientific, Singapore, 1990).

[4] G. Gyorgyi, Phys. Rev. Lett. **64**, 2957 (1990).

[5] W. Krauth, M. Mezard, y J-P. Nadal, Complex Syst. **2**, 387 (1988).

[6] J.A. Hertz, en *Statistical Mechanics of Neural Networks: Proceedings of the Eleventh Sitges Conference*, editado por L. Garrido (Springer, Berlin, 1990).

[7] J. Schrager, T. Hogg, y B.A. Hubermann, Science **242**, 414 (1988).

[8] T. Watkin, A. Rau, y M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[9] F. Rosemblatt, *Principles of Neurodynamics* (Spartan, New York, 1962).

[10] N. Tishby, E. Levin, y S. Solla, en *Proceedings of the International Joint Conference on Neural Networks* (IEEE, New York, 1989), Vol. 2, pág. 403.

[11] E. Levin, N. Tishby, y S. Solla, Proc. IEEE **78**, 1568 (1990).

[12] E. Gardner y B. Derrida, J. Phys. A **22**, 1983 (1989).

[13] L.K. Hansen, R. Pathria, y P. Salamon, J. Phys. A **26**, 63 (1993).

[14] S.F. Edwards y P.W. Anderson, J. Phys. F **5**, 965 (1980).

[15] M. Mezard, G. Parisi, y M.A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).

[16] G. Parisi, J. Phys. A **13**, 1101 (1980).

[17] H.S. Seung, H. Sompolinsky, y N. Tishby, Phys. Rev. A **45**, 6056 (1992); H. Sompolinsky, N. Tishby, y H.S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).

[18] L. Diambra, M.T. Martín, C. Mostaccio y A. Plastino, preprint: La Plata–Th 96/2.

[19] L. Diambra y A. Plastino, aceptado para su publicación en *Phys. Rev. E.*

[20] D. Sherrington y S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).

[21] S. Kirkpatrick y D. Sherrington, Phys. Rev. B **17**, 4384 (1978).

[22] N. Parga y M.A. Virasoro, J. Phys. (Paris) **47**, 1957 (1986)

[23] W. Krauth y M. Mezard, J. Phys. (Paris) **50**, 3057 (1989).

Capítulo 3

Aprendizaje y Teoría de la Información

3.1 Introducción

Como hemos visto en capítulos previos una propiedad muy importante de las redes *feedforward* es su capacidad de generalización o *inferencia.* El "programa" responsable del procesamiento de la información de entrada, subyace en el conjunto de parámetros de la red. En nuestro caso los parámetros de la red están representados por el conjunto pesos sinápticos W_i y umbrales de activación θ, los cuales constituye nuestra hipótesis de trabajo. Cuando esta hipótesis es adecuadamente elegida la red puede mapear los patrones de entrada en la salida correcta. La elección de los pesos sinápticos de la red y los umbrales de activación se vuelve ahora prioritaria. Consecuentemente, muchos de los esfuerzos en este campo han sido dedicados a desarrollar algoritmos de entrenamiento que sean capaces de sintonizar los parámetros de la red de forma tal que esta pueda inferir la respuesta correcta cuando un nuevo patrón de entrada es presentado. Como es lógico, uno desea que estos algoritmos de aprendizaje logren el objetivo en un tiempo computacional prudencial y con una cantidad de ejemplos razonable.

Existen diversos algoritmos para lograr que una red aprenda la regla subyacente a

un conjunto de ejemplos. Básicamente estos esquemas de entrenamiento están basados en la minimización estocástica de una función costo o energía de entrenamiento E_t. El proceso de entrenamiento puede ser visto como un *random walk* sobre la superficie de energía la cual depende del conjunto de ejemplos usados en el entrenamiento [1]. La energía de entrenamiento es, en general una función complicada de los pesos, con una multitud de mínimos locales correspondientes a estados metaestables, como consecuencia de una fuerte frustración en el sistema. Es un hecho bien conocido [2], que en ciertas condiciones el sistema puede quedar atrapado en algunos de estos mínimos espurios de la función energía, con el consecuente empobrecimiento de su capacidad de generalización. Por esta razón, los algoritmos de búsqueda de mínimos incorporan cierto grado de aleatoriedad como un mecanismo de escape de los mínimos no deseados [3]. Entre estos mecanismo encontramos las fluctuaciones térmicas en los procesos de relajación [2] o en simulación de templado [4], o mutaciones en los algoritmos genéticos [5]. Recientemente se ha incorporado también cierto tipo de ruido coloreado a los procesos de relajación [6].

En el marco de los algoritmos de entrenamiento basados en procesos de relajación, el tiempo requerido en remontar la barreras de energías libre es del orden de $t \approx e^{N\Delta f/T}$, donde Δ es la altura de la barrera y T la temperatura del sistema. Por esta razón si existen numerosos estados metaestables la evolución del aprendizaje puede ser anormalmante lenta, quedando el sistema atrapados en algún mínimo local [1]. Esto constituye una seria dificultad si uno desea optimizar el conjunto de parámetros que le otorguen a la red una buena performance en la generalización. Por otro parte, cuando la cantidad de ejemplo P es menor al número de pesos a determinar (es decir cuando $\alpha < 1$) no es posible asegurar que los mínimos de E_t correspondan siempre al conjunto de pesos que otorgan a la red una buena performance en la generalización. En otra palabras, un valor pequeño en la energía de entrenamiento no garantiza una buena generalización. Esto constituye el principal argumento contra las estrategias evolutivas, como los algoritmos genéticos, que se han aplicado con éxito en muchos otros problemas de optimizacion o búsqueda de mínimos.

[1] En ciertas condiciones, puede haber una transición a una fase de vidrio de spin. En esta fase, los estados metaestables están separados por barreras que divergen con N, y las fluctuaciones térmicas están congeladas.

En este capítulo, proponemos una método de entrenamiento para redes de una capa a partir de una herramienta derivada del Principio de Máxima Entropía, en el marco de la Teoría Información (TI) [7]. La descripción de sistemas físicos haciendo uso de la TI ha sido investigada ampliamente [8, 9, 10] a partir de las ideas pioneras de Jaynes, quien propuso una reformulación de la Mecánica Estadística. En los años siguientes, se presentaron extensiones del formalismo para tratar con problemas cuánticos [11, 12].

3.2 Conceptos básicos de la Teoría de la Información

En esta sección presentaremos la forma de caracterizar un sistema físico haciendo uso del Principio de Máxima Entropía de la TI, cuando se supone conocido un conjunto (reducido) de valores medios de observables relevantes del problema. Shannon desarrollo un esquema matemático que permite cuantificar la información contenida en una distribución de probabilidades p_i con $i = 1, ..., M$, definida sobre un conjunto de M sucesos complementarios. Dentro de este esquema, la medida de información faltante o ignorancia esta dada por la expresión:

$$S = -\sum_{i=1}^{M} p_i \ln p_i, \tag{3.1}$$

donde si el logaritmo es en base 2, entonces S esta dado en bits.

Hacia fines de los años 50, Jaynes propuso [8] un metodo de inferencia, conocido como Principio de Maxima Entropía, por medio del cual dio una prescripción para construir una distribución de probabilidades en base al menor número de hipótesis basado en la Teoría de la Información, esta metodología ha sido extensamente aplicada en numeroso problemas de la física y hoy pretendemos usarla a lo largo de esta tesis para desarrollar un método eficiente de entrenamiento de redes neuronales.

Consideremos un sistema X, caracterizado por el conjunto de variables aleatorias A_α con $\alpha = 1, ..., N$, el cual puede estar en el estado i con probabilidad p_i. Asumimos que nuestro conocimiento del sistema esta limitado al conjunto de valores de

expectación

$$\langle A_\alpha \rangle = \sum_{i=1}^{M} p_i A_{\alpha,i} \qquad \alpha = 1, ..., P \quad P < M. \tag{3.2}$$

La cuestión principal consiste en que forma podemos construir la distribución de probabilidades a partir de la informacion contenida en (3.2) de forma tal que pueda predecir los resultados de cualquier medida de los A_α. De todas las distribuciones de probabilidades que reproduzcan los datos conocidos, la que posee mayor poder de inferencia es la que además maximiza nuestra ignorancia o entropía (3.1). Esto nos conduce a un problema de extremalización de la entropía (3.1), sujeta a las ligaduras que imponen las (3.2). Este proceso de maximización con vínculos puede efectuarse mediante la técnica de los multiplicadores de Lagrange

$$\delta_{\{p_i\}} \left[-\sum_{i=1}^{M} p_i \ln p_i - \lambda_0 \left\{ \sum_{i=1}^{M} p_i - 1 \right\} - \sum_{\alpha=1}^{P} \lambda_\alpha \left\{ \sum_{i=1}^{M} p_i A_{\alpha,i} - \langle A_\alpha \rangle \right\} \right] = 0, \tag{3.3}$$

donde λ_0 es el multiplicador que garantiza la normalización, mientras que los λ_α están asociados con los vínculos (3.2). Afortunadamente, el problema variacional puede ser resuelto en forma analítica obteniendo así las p_i

$$p_i = \exp \left[-(1 + \lambda_0) - \sum_{\alpha=1}^{P} \lambda_\alpha A_{\alpha,i} \right]. \tag{3.4}$$

A partir de la condición de normalización, obtenemos

$$\begin{aligned} \lambda_0 &= \ln \textstyle\sum_{i=1}^{M} \exp \left[-\sum_{\alpha=1}^{P} \lambda_\alpha A_{\alpha,i} \right] \\ &= \ln Z(\lambda_1, ..., \lambda_P), \end{aligned} \tag{3.5}$$

donde Z es la función de partición generalizada, introduciendo (3.4) y (3.5) en (3.2), obtenemos un conjunto de P ecuaciones acopladas para determinar los multiplicadores de Lagrange

$$\frac{\partial \ln Z(\lambda_1, ..., \lambda_P)}{\partial \lambda_\alpha} = -\langle A_\alpha \rangle \qquad \alpha = 1, ..., P. \tag{3.6}$$

Resolviendo este sistema de ecuaciones, obtenemos la distribución de probabilidades de Máxima Entropía (3.4). Uno puede probar que esta distribución de probabilidad siempre existe y está univocamente determinada por (3.5) y (3.6) [10].

Frecuentemente, existe disponible una información adicional a (3.2). Uno conoce de antemano que las probabilidades p_i son de la forma

$$p_i = g_{1,i}\, g_{2,i} \tag{3.7}$$

con $g_{1,i}$ conocida y $g_{2,i}$ desconocida. En este caso, la cantidad a maximizar es la llamada entropía relativa, dada por

$$S' = -\sum_{i=1}^{M} p_i \ln\left[p_i/g_{1,i}\right], \tag{3.8}$$

ésta modificación no produce cambios esenciales en (3.4)–(3.4) excepto que (3.4) nos da la distribución desconocida $g_{2,i}$.

En la siguiente sección, investigaremos la aplicación de este formalismo al desarrollo de un método de aprendizaje de una regla. Según este nuevo esquema, el proceso de entrenamiento de la red neuronal es visto como un proceso de inferencia del estado del perceptrón que genera los ejemplos (previamente llamado perceptrón maestro). Primero se construye una distribucion de probabilidades a partir de la información disponible en los ejemplos y seleccionamos luego como nuestra hipótesis de trabajo, aquella configuración de pesos que tenga máxima probabilidad. El concomitante desarrollo constituye uno de los aportes originales de esta Tesis.

3.3 Aprendizaje con Máxima Entropía

En el lenguaje de la Teoría de la Información, se llama nivel de observación [13] al conjunto fijo de observables considerados relevantes para construir el operador estadístico. En el terreno de redes neuronales, los observables están dados por los ejemplos miembros del conjunto de entrenamiento, es decir, la respuestas del sistema a excitaciones conocida. Uno puede usar la información contenida en los ejemplos en formas diferentes con el fin de determinar la distribución de probabilidades de los pesos. Cada una de estas formas conduce a una distribución que pueden exhibir propiedades diferentes. Hasta ahora la elección usual consiste en considerar un único observables, la energía de entrenamiento E_t, obtenida a partir de una expresión que

involucra el conjunto completo de ejemplos [1]. Por lo tanto el nivel de observación esta constituido solo por un observable. Nuestra intención es concentrar los esfuerzos en seleccionar la mejor hipótesis de trabajo, la idea central consiste en trabajar con un nivel de observación que involucre todos los observables disponibles en una forma mas eficiente. Más especificamente, cada miembro del CE sera considerado como un observable.

Hemos considerado por simplicidad el tipo más sencillo de red neuronal *feedforward*, pero los resultados obtenidos pueden ser extendidas en forma sencilla a otros tipos de arquitecturas como las *fully-connected*, o bien a un perceptrón que incorpore acoplamientos de no-lineales (ver próximo capítulo). Sin embargo, la extensión de estas ideas a redes multicapas consiste actualmente un tópico en investigación.

Consideremos un perceptrón con N unidades de entrada S_i conectadas a una unidad de salida ζ cuyo estado está determinado por la ecuación

$$\zeta = g\left(\mathbf{S}\cdot\mathbf{W}-\theta\right) \tag{3.9}$$

donde $g\left(x\right)$ es la función de transferencia de la neurona de salida, la cual es requerida invertible. De esta forma para cada conjunto de pesos $\mathbf{W}$, el perceptrón realiza un mapeo de $\mathbf{S}$ a ζ. En lo que sigue y para simplificar la notación, incluimos el umbral de activación al vector pesos y escribiendo $\mathbf{W}\equiv\left(W_1,...,W_N,\theta\right)$, mientras que los ejemplos estarán dados por $\mathbf{S}\equiv\left(S_1,...,S_N,1\right)$.

Con el fin de seleccionar la hipótesis de trabajo para el perceptrón estudiante, es necesario inferir el estado del perceptrón maestro (PM) a partir del conjunto de entrenamiento $\{\mathbf{S}^\mu,\zeta_0^\mu\}$, con $\mu=1,...,P$, nuestra información disponible. Los patrones de entrada $\mathbf{S}^\mu$ son seleccionados al azar desde el espacio entrada y su correspondiente salida $\zeta_0\left(\mathbf{S}^\mu\right)$ es provista por el PM en el estado $\mathbf{W}_0$ y con función de transferencia g_0.

Consideremos en primer lugar el caso usual en el cual el nivel de observación comprende un único observable, la energía de entrenamiento E_t. Por simplicidad consideraremos la definición más extendida de la energía E_t dada por la desviación

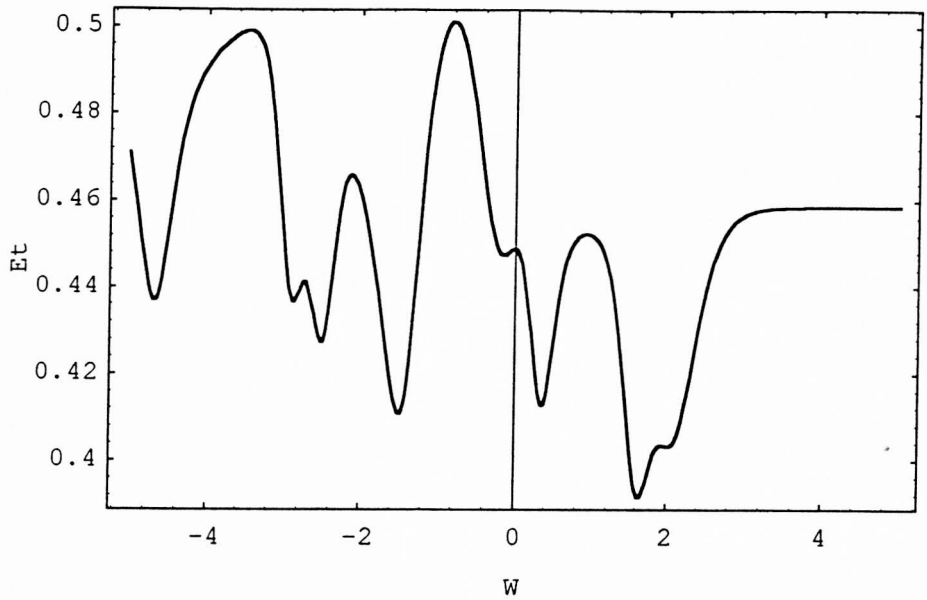

Figura 3.1: Energía de entrenamiento E_t como función de uno de los pesos. Los pesos restantes estan dados al azar.

cuadrática media [14]

$$E_t(\mathbf{W}) = \frac{1}{2}\sum_{\mu=1}^{P}\left(\zeta_0^\mu - g\left(\mathbf{S}^\mu\cdot\mathbf{W}\right)\right)^2. \tag{3.10}$$

Para nuestros propósitos no es importante la forma de E_t, aunque es conveniente aclarar que es posible encontrar otras muchas maneras de definir la función costo. Levin *et al.* [15] mostraron que la distribución estacionaria de los pesos de la red tiene carácter Gibbsiano

$$P(\mathbf{W}) = Z^{-1}\exp\left[-\beta E_t(\mathbf{W})\right], \tag{3.11}$$

donde β es la inversa de la temperatura T y Z la función de partición. La energía de entrenamiento es, en la mayoría de los casos, una función complicada de los pesos $\mathbf{W}$ (ver Fig. 3.1) y no es posible encontrar una expresión general del valor de equilibrio térmico.

Ahora, la idea consiste en introducir una forma más eficiente de aprovechar la información contenida en los ejemplos para obtener una distribución $P(\mathbf{W})$ más suave. En primer lugar escribiremos los ejemplos como

$$g^{-1}\left(\zeta_0^\mu\right) = \mathbf{S}^\mu\cdot\mathbf{W} \tag{3.12}$$

donde $\mathbf{S}^{\mu}$ es la matriz de los patrones de entrada y $g^{-1}(\zeta_0^{\mu})$ es un vector de componentes $\left(g^{-1}(\zeta_0^1), g^{-1}(\zeta_0^2), ..., g^{-1}(\zeta_0^P)\right)$, formado por las salidas. Estos ejemplos constituyen nuestra información acerca del problema y cada uno de los P ejemplos, será considerado como un observable independiente.

Con el fin de determinar los pesos $\mathbf{W}$, introduciremos un procedimiento para construir la distribución de probabilidades de los pesos $P(\mathbf{W})$ compatible con la información disponible [16, 17], basado en el formalismo de Máxima Entropía [7, 8, 9]. Para ello, asumiremos que cada conjunto de pesos del perceptrón estudiante $\mathbf{W}$ tiene probabilidad $P(\mathbf{W})$. Esta distribución está normalizada a la unidad

$$\int P(\mathbf{W}) d\mathbf{W} = 1, \tag{3.13}$$

donde $d\mathbf{W} = dW_1 dW_2 .... dW_N$. Los valores de expectación $\langle W_i \rangle$ están definidos en la forma usual

$$\langle W_i \rangle = \int P(\mathbf{W}) W_i d\mathbf{W}. \tag{3.14}$$

La entropía relativa asociada a la distribución de probabilidades $P(\mathbf{W})$, está definida [7, 8], por la expresión

$$H = -\int P(\mathbf{W}) \ln\left(\frac{P(\mathbf{W})}{P_0(\mathbf{W})}\right) d\mathbf{W}, \tag{3.15}$$

donde $P_0(\mathbf{W})$ es la distribución de probabilidades *a priori* [7, 8, 9]. La elección de $P_0(\mathbf{W})$ depende de nuestro conocimiento *a priori* sobre la arquitectura del PE y no de los ejemplos. Cada uno los P ejemplos es considerados en forma individual y reescritos en la forma

$$g^{-1}(\zeta_0^{\mu}) = \mathbf{S}^{\mu} \cdot \langle \mathbf{W} \rangle . \tag{3.16}$$

Nuestro propósito es ahora maximizar la entropía (3.15), sujeta a las restricciones impuestas por nuestro nivel de observación (3.16). De estas forma, la expresión a maximizar se puede escribir como

$$H' = -\int \left\{ P(\mathbf{W}) \ln\left[\frac{P(\mathbf{W})}{P_0(\mathbf{W})}\right] + \lambda_0 P(\mathbf{W}) + (\mathbf{S}^{\mu})^t \overrightarrow{\lambda} \mathbf{W} P(\mathbf{W}) \right\} d\mathbf{W} \tag{3.17}$$

donde λ_0 y $\overrightarrow{\lambda}$ son los multiplicadores de Lagrange asociados con la condición de normalización (3.13) y con las restricciones (3.16), respectivamente. Calculando la

variación de H' con respecto a $P(\mathbf{W})$ e igualando a cero, obtenemos una expresión para la distribución de probabilidad, dada por

$$P(\mathbf{W}) = \exp\left[-(1+\lambda_0) - \mathbf{\Gamma}\cdot\mathbf{W}\right] P_0(\mathbf{W}), \tag{3.18}$$

donde $\mathbf{\Gamma} = (\mathbf{S}^\mu)^t \overrightarrow{\lambda}$. Esta es la distribución de probabilidades *a posteriori* obtenida por Máxima Entropía. Los multiplicadores de Lagrange deben ser determinados en forma autoconsistente a partir de (3.16) después de una adecuada elección de $P_0(\mathbf{W})$.

A la hora de elegir adecuadamente P_0, existen algunas consideraciones que deben ser tenidas en cuenta. Muchas veces, en la implementación práctica de una red neuronal es necesario hacer algunas concesiones a las posibilidades técnica. Tecnológicamente, es más sencillo construir una red cuyos pesos solo toman dos valores (pesos binarios) que una red, en la cual estos pueden tomar cualquier valor real (pesos reales). La elección de la *a priori* distribución nos ofrece una alternativa de incorporar información adicional a los ejemplos, ya sea porque conocemos la arquitectura del perceptrón, en lo concerniente a los pesos; o porque restringimos el espacio de los pesos, por ejemplos exigir que sean binarios. En este capítulo, analizaremos algunos de estos problemas con diferentes elecciones de P_0.

3.3.1 Perceptrón con pesos reales

En primer lugar consideremos el problema en el cual no existe ninguna restricción sobre los pesos que pueda tomar el PE y que no conocemos nada acerca de los pesos del PM. En este caso la elección más sencilla [16, 17] consiste en tomar P_0 proporcional a $\exp(-\mathbf{W}.\mathbf{W}/2a)$, con un parámetro libre a. Reemplazando ésta distribución *a priori* en (3.18), se obtiene una forma Gaussiana para la distribución de probabilidades, centrada en $\langle W_i\rangle = -a\Gamma_i$.

$$P(\mathbf{W}) = \frac{1}{(2\pi a)^{N/2}} \exp\left[-\frac{1}{2a}(\mathbf{W} - a\mathbf{\Gamma})^2\right]. \tag{3.19}$$

Con esta elección de P_0, la distribución de probabilidades para cualquier g invertible es Gaussiana en cada W_i, es decir

$$P(\mathbf{W}) = Z^{-1} \prod_i^N \exp\left[-\beta\left(W_i - \langle W_i\rangle\right)^2\right]. \tag{3.20}$$

Como podemos ver el parámetro $2a$ juega el rol de la temperatura T. Esta distribución de probabilidades tiene forma Gibbsiana con una energía que tiene un único mínimo en $W_i = -a\Gamma$. Esta ventaja se conserva aún cuando se trate de un problema en el cual el TM y PE no tienen la misma función de transferencia y el problema no es aprendible.

A partir de la definición de $\mathbf{\Gamma}$ y de las restricciones (3.16) se pueden eliminar los multiplicadores de Lagrange $\overrightarrow{\lambda}$ y obtener una expresión para $\langle W_i \rangle$ en términos de nuestra información disponible, los ejemplos

$$\langle \mathbf{W} \rangle = I_{ps} \left[\mathbf{S}^{\mu} \right] g^{-1} \left(\zeta_0^{\mu} \right), \tag{3.21}$$

donde $I_{ps} \left[\mathbf{S}^{\mu} \right] = \left(\mathbf{S}^{\mu} \right)^t \left(\mathbf{S}^{\mu} \left(\mathbf{S}^{\mu} \right)^t \right)^{-1}$es la pseudoinversa de Moore-Penrose. Este es un resultado es similar a la regla de Personnaz *et al.* [18], utilizada en almacenamiento de patrones.

Ahora elegimos como hipótesis de trabajo, la configuración de pesos más probable. En este caso la configuración más probable coincide con el valor medio y proponemos como regla de aprendizaje tomar

$$\mathbf{W} = I_{ps} \left[\mathbf{S}^{\mu} \right] g^{-1} \left(\zeta_0^{\mu} \right). \tag{3.22}$$

Es fácil darse cuenta que con ésta regla de aprendizaje el error de entrenamiento es cero para todo α, si la regla es aprendible.

Como una forma de comparar los dos niveles de observación considerados aquí podemos estudiar las superficie de energías de las distribuciones (3.19) y (3.11) por simulaciones con Montecarlo. A partir de una configuración inicial $\mathbf{W}$, generamos por medio una pequeña modificacione un nueva configuración levemente diferente $\mathbf{W}'$, calculamos el cambio en la energía $\Delta E = E\left(\mathbf{W}'\right) - E\left(\mathbf{W}\right)$. Si esta diferencia es positiva el cambio es rechazado. En cambio si ΔE es negativo el cambio es aceptado con probabilidad $\exp(-\beta \Delta E)$. Este paso es repetido muchas veces. Este simple procedimiento nos permite conocer algunos detalles concernientes a la naturaleza de la rugosidad de la superficie de energía. En la Fig. 3.2 ilustramos las dos situaciones aquí tratadas.

En el caso del nivel de observación trivial (E_t como ligadura), el paisaje de energía posee muchos estados metaestables, como resultado de una gran frustración. Como el

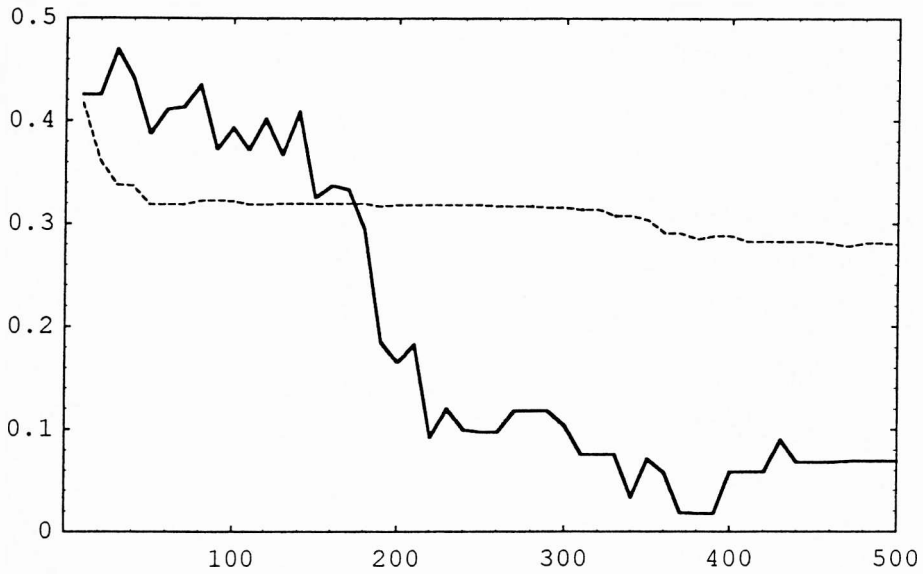

Figura 3.2: Energía de entrenamiento E_t según el número de iteraciones para los dos tipo de nivel de observación. La línea llena corresponde a nuestro esquema y la línea de trazo al nivel de observación trivial.

proceso de optimización debe ser llevado a cabo a baja temperatura, el sistema quedar atrapado en mínimos locales con la consecuente pobre performance en la generalización. Con la propuesta de considerar a cada uno de los P ejemplo como observables independiente, uno obtiene una superficie de energía más suave con un único mínimo global y el sistema converge rápidamente a él. El mismo comportamiento es observado en el caso en que la función de transferencia del PM y PE son diferentes. En este caso, debido a que el problema es no-aprendible el E_t no se anula.

3.3.2 Perceptrón de pesos binarios

Que ocurre si el PE debe tener pesos binarios? El presente formalismo puede ser utilizado para este problema. En este caso, la elección más acertada de la *a priori* distribución P_0, está dada por

$$P_0 = \prod_i^N \left\{ \exp\left[-\frac{(W_i - 1)^2}{2a} \right] + \exp\left[-\frac{(W_i + 1)^2}{2a} \right] \right\}. \tag{3.23}$$

Esta distribución *a priori*, constituye un forma *suave* de la restricción sobre el espacio de pesos [19]. En el límite $a \to 0$, los pesos se convierten en variables binarias

y solo pueden tomar ± 1. Usando (3.18) y (3.23) obtenemos una nueva distribución de probabilidades con función de partición dada por

$$Z = (2\pi a)^{\frac{N}{2}} \prod_i^N \exp\left(\frac{a}{2}\Gamma_i^2\right) 2\cosh(\Gamma_i). \tag{3.24}$$

El cálculo de los valores de expectación conduce a

$$\langle W_i \rangle = -a\Gamma_i - \tanh(\Gamma_i). \tag{3.25}$$

A diferencia del caso previo con pesos reales, no es posible la eliminación de los multiplicadores de Lagrange λ_i, pero si obtenemos N ecuaciones trascendentes para los Γ_i

$$I_{ps}[\mathbf{S}^\mu] g^{-1}(\zeta_0^\mu) + a\boldsymbol{\Gamma} + \tanh(\boldsymbol{\Gamma}) = 0 \tag{3.26}$$

donde $\tanh(\boldsymbol{\Gamma}) \equiv (\tanh\Gamma_1, ..., \tanh\Gamma_N)$. Por esta razón no es posibles encontrar una expresión cerrada para la distribución de probabilidades expresada en términos del conjunto de ejemplos (3.12). La nueva distribución de probabilidades puede ser expresada como una suma de dos Gaussianas, pesadas con $p_i^{\pm} = \exp(\pm\Gamma_i)/2\cosh(\Gamma_i)$,

$$\begin{aligned} P(\mathbf{W}) &= \frac{1}{(2\pi a)^{N/2}} \prod_i^N \left\{ p_i^+ \exp\left[-\frac{1}{2a}(W_i + a\Gamma_i + 1)^2\right] \right. \\ &+ \left. p_i^- \exp\left[-\frac{1}{2a}(W_i + a\Gamma_i - 1)^2\right] \right\}. \end{aligned} \tag{3.27}$$

Esta expresión, no tiene la forma Gibbsiana $Z^{-1}\exp(-\beta E)$, y no es posible considerar al parámetro a asociado con la temperatura de la misma forma que en el caso previo, sino más bien como una medida de la suavidad con que imponemos la restricción binaria. Ahora nuestro mayor es interés considerar el límite en el cual los pesos solo pueden tomar valores ± 1. Tomando el límite $a \to 0$ en (3.25), obtenemos una expresión analítica para los multiplicadores de Lagrange y pueden ser eliminados

$$\boldsymbol{\Gamma} = -\tanh^{-1}\left(I_{ps}[\mathbf{S}^\mu] g^{-1}(\zeta_0^\mu)\right). \tag{3.28}$$

En este límite, la distribución de probabilidades (3.27) se puede escribir como

$$P(\mathbf{W}) = \prod_i^N \left\{ p_i^+ \, \delta(W_i + 1) + p_i^- \, \delta(W_i - 1) \right\}, \tag{3.29}$$

Tabla 3.I: Probabilidades inferidas (p_i^+) para algunos pesos a partir de p ejemplos para una perceptrón con $g_0(x) = g(x) = \tanh(x)$ y $N = 30$.

W_0	$p=3$	$p=6$	$p=12$	$p=21$	$p=27$	$p=30$
1	0.5910	0.9711	0.7795	0.6949	0.8710	1.
-1	0.5189	0.6428	0.7130	0.2302	0.0117	0.
-1	0.2945	0.2761	0.3053	0.1159	0.1837	0.
1	0.7993	0.5799	0.5843	0.7991	0.9827	1.
1	0.5098	0.8356	0.8528	0.8193	0.8337	1.

donde los coeficientes $p_i^{\pm}$ son ahora las probabilidades que el *i-ésimo* peso W_i tome el valor ± 1 respectivamente. Estas probabilidades tienen una expresión analítica dada solamente en términos del conjunto de ejemplos (3.12),

$$p_i^{\pm} = \frac{\exp\left[\mp \tanh^{-1}\left[\{I_{ps}\left[\mathbf{S}^{\mu}\right] g^{-1}\left(\zeta_0^{\mu}\right)\}_i\right]\right]}{2\cosh\left[\tanh^{-1}\left[\{I_{ps}\left[\mathbf{S}^{\mu}\right] g^{-1}\left(\zeta_0^{\mu}\right)\}_i\right]\right]}. \tag{3.30}$$

Este resultado no depende de la arquitectura de PM sino solamente de los ejemplos y de la función de transferencia del PE. En la Tabla 3.I, podemos ver las probabilidades inferidas de algunos pesos para diferentes valores del parámetro α.

La hipótesis de trabajo puede ser seleccionada ahora, maximizando (3.29); es decir, tomamos $\mathbf{W}$ tal que si $p_i^+ > p_i^-$ (o $p_i^+ < p_i^-$) entonces $W_i = 1$ (o $W_i = -1$), como hipótesis de trabajo. Esta receta, pude ser implementada sencillamente haciendo

$$\begin{aligned} W_i &= \text{sign}\left[p_i^+ - p_i^-\right] \\ &= \text{sign}\left[\{I_{ps}\left[\mathbf{S}^{\mu}\right] g^{-1}\left(\zeta_0^{\mu}\right)\}_i\right]. \end{aligned} \tag{3.31}$$

3.3.3 Un mecanismo iterativo

Como hemos visto, si la regla es aprendible, solo es suficiente tener N ejemplos ($\alpha = 1$) para inferir el estado correcto $\mathbf{W}_0$. Muchas veces, como discuturemos más adelante, la regla no es exactamente aprendible, en ese caso logramos la mejor performance cuanto mayor sea el número de ejemplos disponibles. A partir de las expreciones (3.22) y (3.31), podemos apreciar, que el tiempo de cómputo de los pesos, crece rápidamente con la cantidad de ejemplos, debido a que tenemos que diagonalizar una matriz de $P \times P$. Por lo tanto, a la hora de trabajar con un gran conjunto de ejemplos, conviene tener en cuenta la alternativa que nos brinda la libertad de elección de P_0.

La idea consiste en particionar el CE y considerar una porción Π_1. Según vimos en las secciones previas (3.16–3.18), este subconjuto tiene asociada una distribución de Maxíma Entropía. Esta distribución es ahora nuestra información *a priori* y puede tomarse como la nueva P_0 para determinar otra distrbución de probabilidades asociada con la siguiente porcion de ejemplos Π_2. Como sabemos nuestro esquema prevee (3.18), donde ahora $P_0(\mathbf{W}) = \exp\left[-1/2\,(\mathbf{W} - \langle\mathbf{W}\rangle_1)^2\right]$. De esta manera obtenemos una Gaussiana centrada en

$$\langle\mathbf{W}\rangle_2 = -a\boldsymbol{\Gamma} + \langle\mathbf{W}\rangle_1 . \tag{3.32}$$

Teniendo en cuenta la definición de Γ y los (3.16), podemos eliminar los multiplicadores, obteniendo

$$\langle\mathbf{W}\rangle_2 = \langle\mathbf{W}\rangle_1 + I_{ps}\left[\mathbf{S}^\mu\right]\left[g^{-1}\left(\zeta_0^\mu\right) - \mathbf{S}^\mu \cdot \langle\mathbf{W}\rangle_1\right], \tag{3.33}$$

donde los subíndices 1 y 2 se refieren a los subconjuntos Π_1 y Π_2 respectivamente. Como en las secciones previas, nuestra hipotesis de trabajo estara dada por los pesos más probables compatibles con la información disponible. En este caso coincide con el valor medio (3.33). De esta manera, hemos obtenido un camino iterativo, como la regla de Hebb, para encontrar los pesos más adecuados al CE

$$\mathbf{W}_{new} = \mathbf{W}_{old} + I_{ps}\left[\mathbf{S}^\mu\right]\left[g^{-1}\left(\zeta_0^\mu\right) - \mathbf{S}^\mu \cdot \mathbf{W}_{old}\right]. \tag{3.34}$$

3.4 Análisis de la performance

Si uno desea evaluar la performance del algoritmo debe estudiar la evolución tanto del error de generalización ϵ_g como el error de entrenamiento promedio. Estas cantidades cuantifican la bondad de la red con respecto al espacio de ejemplos y con respecto al conjunto de entrenamiento, respectivamente. Para evaluar la performance de generalización en la cual estamos interesados, definimos el error de generalización en términos del promedio sobre todo el espacio de las entradas, de la distancia entre la salida deseada ζ_0 y la salida de la red ζ correspondiente a una dada señal de entrada $\mathbf{S}$

$$\epsilon_g(\mathbf{W}) = \frac{1}{2}\int d\mu(\mathbf{S})\left[g_0(\mathbf{W}_0{\cdot}\mathbf{S}) - g(\mathbf{W}\cdot\mathbf{S})\right]^2 \tag{3.35}$$

Según el cálculo realizado en el Apéndice A, el error de generalización está dado por

$$\epsilon_g(R) = \frac{1}{4\pi} \int \frac{dxdy}{\sqrt{1-R^2}} \exp\left[\frac{x^2+y^2-2xyR}{2(1-R^2)}\right] [g_0(x) - g(y)]^2 \tag{3.36}$$

donde $R = N^{-1}\mathbf{W}_0 \cdot \mathbf{W}$ es el solape entre los vectores pesos del PM y PE. El comportamiento del error de generalización está determinado por el parámetro R.

En ésta sección estudiaremos la performance de un perceptrón entrenado con este esquema en diferentes casos. Para ello calculamos el valor promedio de R sobre 200 muestras en una red con $N = 80$ para diferentes valores de α. Obtenemos la curva de aprendizaje correspondiente al error de generalización evaluando (A.5) para diferentes valores de R. En la Fig. 3.3 podemos ver el error de generalización en la situación en que P_0 es Gaussiana. En este caso, los pesos están dados por (3.21). La función de transferencia tanto del PE como del PM son idénticas, $g_0(x) = g(x) = \tanh(x)$ (línea llena). El error de generalización se anula en $\alpha = 1$. La curva de línea de puntos corresponde al caso en el cual las funciones de transferencias son diferentes. En este caso la regla no es aprendible, es decir que el error de generalización no se anula, sino que disminuye hasta alcanzar un mínimo $\epsilon_{\min}$ el cual depende de las funciones de transferencia concomitantes, si bien R alcanza la unidad para $\alpha = 1$.

Por otro lado, es interesante estudiar el problema en el cual los pesos sinápticos de PM son binarios y la distribución *a priori* esta dada por (3.23) en el límite $a \to 0$. En este caso, el conjunto de pesos $\mathbf{W}$ más probables compatibles con los ejemplos es dado por la expresión (3.31). En la Fig 3.3 podemos ver en línea punteada, que el ϵ_g exhibe una saturación cerca de $\alpha = 1$, este efecto, podría ser una consecuencia de la información adicional introducida en P_0. El hecho importante a destacar en este nuevo esquema de entrenamieto para redes de pesos binarios, es la ausencia de transiciones de fase. En el capítulo previo vimos que la transición de fase de un estado generalización pobre, a una estado de generalización perfecta (con $R = 1$) en un perceptrón de salida booleana, esta transiciones están presentes aún cuando la salida es lineal [20], siempre que los pesos sean sean discretos. Con nuestro esquema, la generalización perfecta se alcanza en $\alpha = 1$.

Como otro ejemplo más de la aplicación de nuestro método, consideraremos el caso de una regla no-aprendible, es posible concebir que un PE con pesos binarios

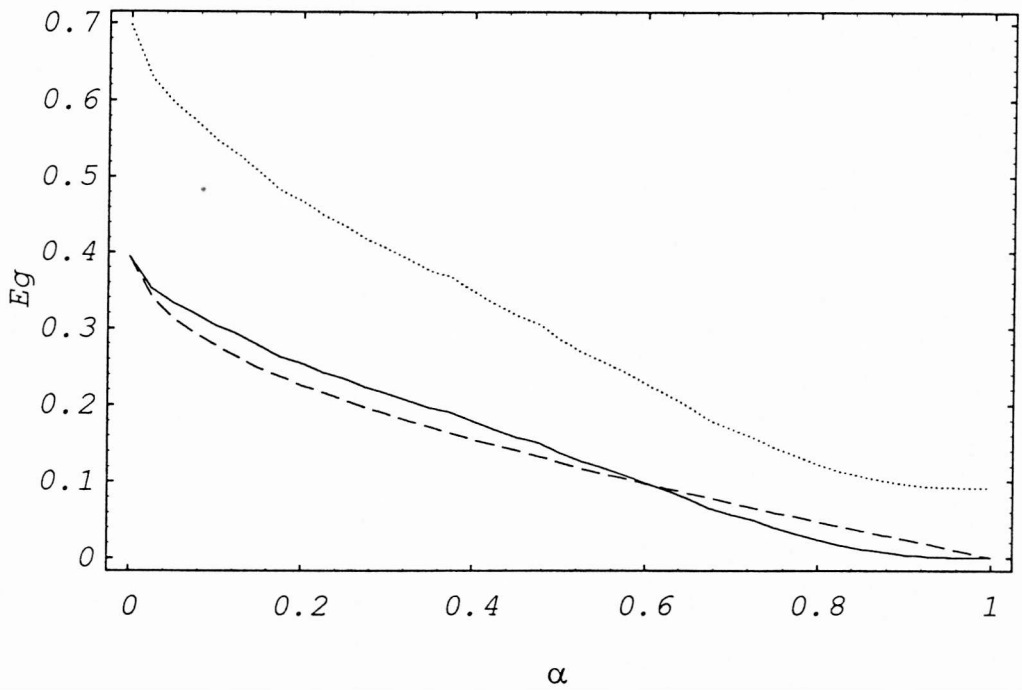

Figura 3.3: Error de Generalización versus α: en el caso aprendible con $g_0(x) = g(x) = \tanh(x)$ y (línea llena), y no aprendible con $g_0(x) = x$ and $g(x) = \tanh(x)$ (línea punteada). En ambos casos, P_0 corresponde a pesos binarios. En línea de trazos tenemos el problema aprendible con P_0 Gaussiana. Hemos tomado $N = 80$ y promediamos sobre 200 casos.

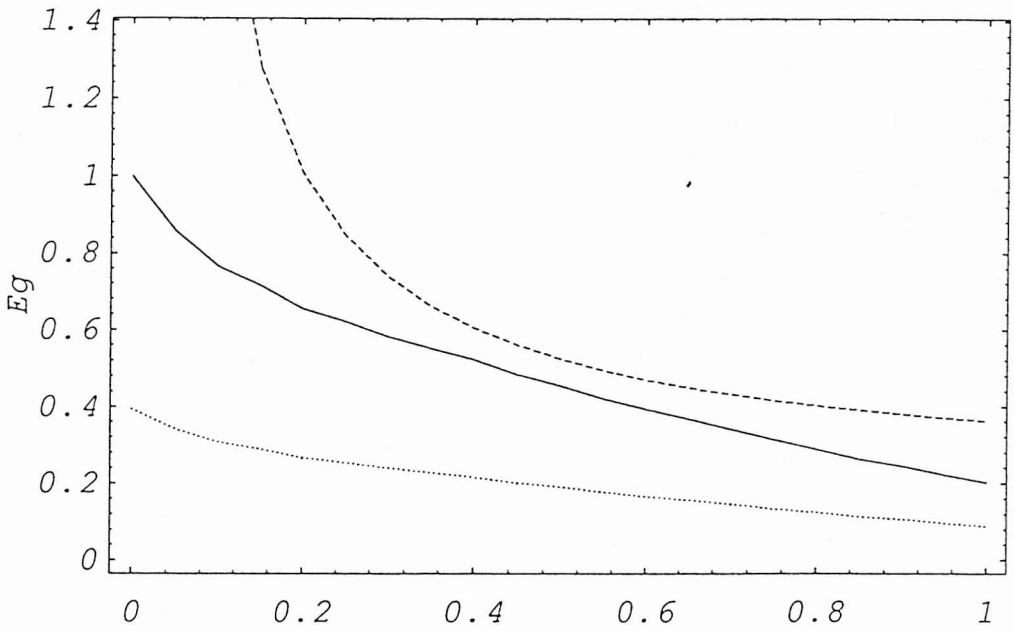

Figura 3.4: Error de generalización versus α en un problema no aprendible: la línea llena corresponde a nuestro método y la línea de trazos a la solución de simetría de réplica. con $g_0(x) = g(x) = x$ en ambos casos. La línea punteada corresponde a $g_0(x) = x$ and $g(x) = \tanh(x)$. Otra vez, $N = 80$ y promediamos sobre 200 casos.

pueda aprender, no sin libre de error, la regla que subyace a los ejemplos generados por un PM con pesos reales (mismatched weights). Por simplicidad usamos funciones de transferencia lineales tanto para PM como para PE. Para una distribución de las componentes de $\mathbf{W}_0$ Gaussiana, el overlap máximo $R_{\max}$ es obtenido por $\mathbf{W} = \text{sign}[\mathbf{W}_0]$. En el límite termodinámico $N \to \infty$ $R_{\max} = \sqrt{2/\pi}$. La solución de simetría de réplica [2] da un comportamiento asintótico con α, para el error de generalización, dado por

$$\epsilon_g = \epsilon_{\min} + \frac{\epsilon_{\min} R_{\max}}{\alpha} + O\left(\alpha^{-2}\right) \tag{3.37}$$

con $\epsilon_{\min} = 1 - R_{\max} = 0.202$. Nuestro esquema muestra un escenario totalmente diferente. En la Fig 3.4 podemos ver el hecho que $\epsilon_g = \epsilon_{\min}$ para $\alpha = 1$.

Bibliografía

[1] T. Watkin, A. Rau, y M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[2] H.S. Seung, H. Sompolinsky, y N. Tishby, Phys. Rev. A **45**, 6056 (1992).

[3] D.E. Rumelhart y J.L. McClelland, *Parallel Distributed Processing*, (MIT, Cambridge, MA., 1986).

[4] S. Kirkpatrick, C. Gellat, y M. Vecci, Science **220**, 671 (1983).

[5] J. Holland, *Evolution, Learning and Cognition*, editado por Y.S. Lee (World Scientific, Singapore, 1988).

[6] Y. Hayakawa, A. Marumoto, and Y. Sawada, Phys. Rev. E **51**, R2693 (1995).

[7] C.E. Shannon y W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Chicago, Ill), 1949.

[8] E.T. Jaynes, Phys. Rev. **106**, 620 (1957); Phys. Rev. **108**, 171 (1957).

[9] R.D. Levine y M. Tribus, *The Maximum Entropy Principle* (MIT Press, Boston MA), 1978.

[10] A. Katz, *Principle of Statistical Mechanics* (Freeman, San Fransico, 1967).

[11] N. Canosa, A. Plastino y R. Rossignoli, Phys. Rev. A **40**, 519 (1989).

[12] N. Canosa, R. Rossignoli y L. Diambra, Phys. Lett. A **185**, 133 (1994).

[13] E. Fick y G. Sauermann, *Quantenstatistik dinamischer Prozesse, Band I* (Verlag Harri Deutsch, Francfurt/Main), 1983.

[14] R. Meir y J.F. Fontanari, Phys. Rev. A **45**, 8874 (1992).

[15] E. Levin, N. Tishby y S. Solla, Proc. IEEE **78**, 1574 (1990).

[16] L. Diambra, J. Fernandez, y A. Plastino, Phys. Rev. E **52**, 2887 (1995).

[17] L. Diambra y A. Plastino, aceptado para su publicación en Phys. Rev. E.

[18] L. Personnaz, I. Guyon y G. Dreyfus, J. Physique Lett. **16,** L359 (1985); Phys. Rev. A **34**, 4217 (1986).

[19] H. Sompolinsky, N. Tishby, y H.S. Seung, Phys. Rev. Lett. **65**, 1683 (1990).

[20] L. Diambra, C. Mostaccio, M.T. Martín, y A. Plastino, preprint: La Plata–Th 96/2.

Capítulo 4

Aplicaciones a Problemas de Física

4.1 Introducción

Las redes neuronales y las estrategias adaptativas en general, han mostrado una gran versatilidad en la resolución de problemas computacionales en los cuales, los métodos más ortodoxos han fracasado. En los últimos años, han surgido numerosas aplicaciones de redes neuronales y de las estrategias evolutivas, tanto a problemas científicos como tecnológicos en una gran variedad de campos.

-) Identificación y clasificación de objetos físicos.
-) Predicción de series temporales caóticas.
-) Reconstrucción de imágenes distorsionadas por ruido.
-) Predicción de estructura secundaria de proteínas.
-) Reconocimiento de fonemas.

Cada arquitectura con sus correspondiente protocolo de aprendizaje, presenta una habilidad intrínseca en la resolución de un problema particular. Así como las redes

fully-connected han sido utilizadas con éxito en problemas de reconstrucción de patrones, los algoritmos de aprendizaje no-supervisados en mapas topológico tienen utilidad en la categorización de datos. Sin embargo, la mayoría de las aplicaciones a problemas científicos, involucran aprendizaje supervisado de redes multicapas *feed-forward* con unidades analógicas.

En este capítulo, mostraremos como es posible aplicar redes neuronales *feedforward* a algunos problemas de física, como ser la reconstrucción de funciones de onda a partir de valores de expectación y la predicción de series temporales caóticas, e inclusive en el ajuste de de rectas. En estos caso, las redes son usadas para construir un modelo fenomenológico de estos sistemas, para la subsecuente predicción. Dada un conjunto representativo de ejemplos, juntos con una regla efectiva de aprendizaje, tales redes son capaces de capturar las correlaciones físicas esenciales que gobiernan las asociaciones entrada–salida, que pertenecen al conjunto de ejemplos y de esta manera, *predecir* la salida correcta cunado una nueva señales de entrada es presentada. Esta predicción es hecha sobre la base de una hipótesis de trabajo, en nuestro caso representada por los pesos sinápticos. Nuestra atención en este capítilo, está dirigida a la implementación del protocolo de aprendizaje desarrollado en el capítulo previo, a los problemas mencionados.

4.2 Arquitectura y entrenamiento de la red

En las redes *feedforward* multicapas, los valores de un conjunto apropiado de variables están codificados en los patrones de entrada como el estado de las unidades de entrada, la información contenida en estos patrones es analizados y procesada por una o más capas de neuronas intermedias, dependiendo de la complejidad del problema [1, 2, 3, 4, 5]. La señal de salida obtenida en la última capa de la red neuronal, da en forma apropiadamente codificada los resultados del procesamiento de la red, ya sea la versión de una imagen completa, o el cómputo de las cantidades físicas requeridas.

El enorme poder de procesamiento de las redes multicapas está basado en la alta nolinealidad del mapeo entre las unidades de entradas y las de salidas. El método de entrenamiento más conocido y usado para este tipo de redes es un algoritmo de

gradiente descendente generalizado, conocido como retropropagación del error [2], en el cual los pesos son modificados en forma iterativa de forma tal que disminuya la diferencia entre la salida actual de la red y la salida deseada. Este paso debe ser aplicado muchas veces hasta alcanzar que la salida de la red se aproxime a la deseada, tanto como el problema lo requiera. Esto significa disminuir la función energía costo, tanto como sea posible. Sin embargo, muchas veces una buena performance de la red en los patrones de entrada no garantiza la generalización, el aprendizaje de la regla. La bondad de la red neuronal en la generalización y los recursos computacionales [1] requeridos en la etapa de entrenamiento, depende fuertemente de algunos factores que hacen a la arquitectura de la red, como ser el número de capas intermedias y del número de neuronas en cada capa. Si existen pocos neuronas intermedias es posible que la red, después de un largo proceso de entrenamiento, no pueda suministrar el mapeo adecuado y ser incapaz de generalizar correctamente. Por otro lado, si el número de neuronas intermedias es muy grande, existen una gran cantidad de soluciones insatisfactorias a la hora de generalizar, a las cuales el algoritmo puede converger. Estas dificultades constituyen una gran desventaja a la hora de seleccionar el esquema de trabajo y pueden acarrearnos tediosos proceso de construcción y aprendizaje de la red neuronal.

Básicamente los problemas a tratar aquí, consisten en la representación de una función continua, la cual es conocidas solo en un conjunto discretos de puntos. Como un primer ejemplos consideraremos un ejemplos en el cual, la arquitectura de un perceptrón con una solo capa, es suficiente. Sin embargo, muchas veces encontramos que en estos tipos problemas, la simple arquitectura del perceptrón no basta para procesar la información de la señal de entrada y es necesario considerar una arquitectura más complicada.

[1] Cantidad de ejemplos y velocidad de aprendizaje de los ejemplos

Tabla 4.I: Error cuadrático medio promediado sobre 3000 casos. P es el número de ejemplos de entrenamiento y R indica el intervalo donde fue evaluada la red. $g(x) = x$.

R	$P = 1$	$P = 2$ (x10^{-6})	$P = 5$ (x10^{-8})	$P = 8$ (x10^{-9})	$P = 9$ (x10^{-10})	$P = 10$ (x10^{-30})
[-1,1]	0.165	1.02	1.85	2.64	1.30	5.08
[-5,5]	4.08	140	69.3	40	94.3	152
[-50,50]	422	1561	1284	105	527	17434

4.2.1 Una aplicación simple

Las redes *feedforward* con capas intermedias pueden representar cualquier función suave y continua de $\Re^n \rightarrow \Re^m$. Como ilustración del método de aprendizaje consideraremos un problema simple en el cual no se requiere capas intermedias. La tarea en aprender consiste en hallar los coeficientes de una recta que ajusta diez puntos experimentales [9]. Cada ejemplo usados para el entrenamiento de la red consiste en un conjunto de diez puntos de la recta con abcisas fijas, como input y los valores obtenidos por cuadrados mínimos como salida. Estos ejemplos están restringidos en un cierto intervalo Δ. Este problema constituye una regla exactamente aprendible si la función de transferencia es lineal. Para una red con una función de transferencia nolineal $g(x)$ la representación no sera exacta. Estudiaremos la performance de una red en los dos casos, es decir con $g(x) = x$ y con $g(x) = \tanh(x)$.

En la Tabla 4.I, mostramos el error de generalización según el número de ejemplos presentados, cuando la $g(x) = x$. Observamos también que la hipótesis de trabajo obtenidad por nuestro método, no solo muestra un error despreciable en el intervalo en que la red fue entrenada, sino también más allá de este intervalo. Esto refleja el hecho, que en un problema exactamente aprendible el método es capaz de hallar los pesos exactos.

Esta no es la situación encontrada cuando $g(x)$ es no lineal (Tabla 4.II). La red solo tienen buena performance en el intervalo donde fue entrenada, como se puede apreciar en la Fig. 4.1, donde hemos graficado las curvas de nivel del error generalización en función de los coeficientes de la recta, los puntos negros indican la posición de los ejemplos con que la red ha, sido entrenada. Sus ubicaciones corresponden a valles en la superficie de error.

Tabla 4.II: Error cuadrático medio promediado sobre 3000 casos. P es el número de ejemplos de entrenamiento. $g(x) = \tanh(x)$.

$P = 1$	$P = 2$	$P = 4$	$P = 5$	$P = 7$	$P = 10$
0.164	5.3110^{-3}	3.9810^{-4}	5.8110^{-4}	4.9910^{-4}	1.7410^{-3}

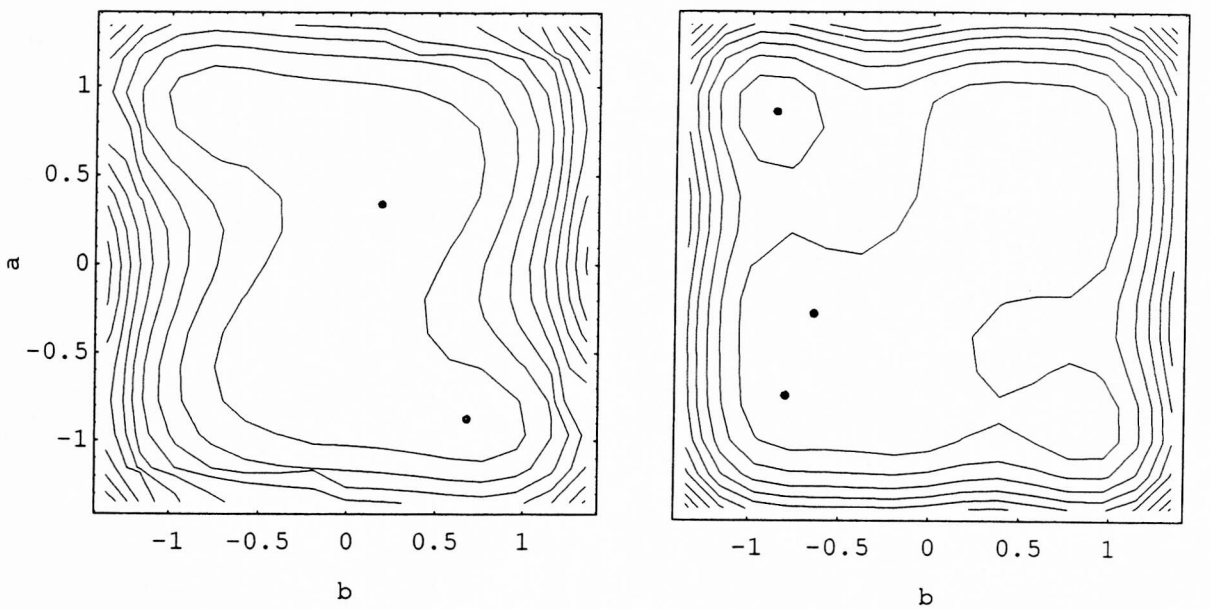

Figura 4.1: Curvas de nivel para el error de generalización como una función de los coeficientes a y b de la recta $ax+b$. La función de transferencia usada es $g(x) = \tanh(x)$. Los puntos negros representan, las ejemplos usados en el proceso de entrenamiento.

Otra característica notable es la pequeña cantidad de ejemplos que son necesarios para lograr una buena generalización. Estas es diferencias notables con respecto de los métodos más ortodoxos en los cuales a menor cantidad de ejemplos la superficie de energía es más rugosa y la convergencia se hace muy lenta, cuando no queda atrapado en mínimos locales.

4.3 Implementación de una red más compleja

Por las razones expuestas arriba, es conveniente explorar otros tipos arquitectura y esquema de aprendizaje, más efectivos tanto en la performance como en los costos

computacionales. En muchos casos [6, 7], podemos lograr un mapeo tan aproximado como uno quiera, por medio de una red tipo perceptrón pero con neuronas cuyo potencial de membrana tiene una dependencia nolineal de los estados de las neuronas de entrada. Consideremos una red con N unidades de entrada ξ_i conectadas a una neurona de salida ζ cuyo estado esta determinado por $\zeta = g(h)$, donde g es la función de transferencia y h el potencial de membrana el cual depende del estado de las neuronas de la capa de entrada $h = h(\xi_1, \xi_2, ..., \xi_N)$[8]. En principio, h es completamente general y puede tener cualquier dependencia en ξ_i. Para nuestros fines prácticos, consideremos una expansión de h como serie de potencia alrededor de ξ_i^0.

$$h = h^0 + \sum_i^N \frac{\partial h^0}{\partial \xi_i}(\xi_i - \xi_i^0) + \frac{1}{2}\sum_{ij}^N \frac{\partial^2 h^0}{\partial \xi_i \partial \xi_j}(\xi_i - \xi_i^0)(\xi_j - \xi_j^0) + \tag{4.1}$$
$$+ \frac{1}{6}\sum_{ijk}^N \frac{\partial^3 h^0}{\partial \xi_i \partial \xi_j \partial \xi_k}(\xi_i - \xi_i^0)(\xi_j - \xi_j^0)(\xi_k - \xi_k^0) + ...$$

Esta expansión puede ser reescrita en una forma mas familiar, que admite una interpretación más sencilla

$$h = \theta + \sum_i w_i \xi_i + \sum_{ij} w_{ij}\xi_i\xi_j + \sum_{ijk} w_{ijk}\xi_i\xi_j\xi_k + .. \tag{4.2}$$

donde el umbral esta definido ahora por $\theta = h^0 + \sum_i^N \frac{\partial h^0}{\partial \xi_i}\xi_i^0 + \frac{1}{2}\sum_{ij}^N \frac{\partial^2 h^0}{\partial \xi_i \partial \xi_j}\xi_i^0\xi_j^0 + .. $, los pesos lineales $w_i = \sum_i^N \frac{\partial h^0}{\partial \xi_i} + \sum_j^N \frac{\partial^2 h^0}{\partial \xi_i \partial \xi_j}\xi_j^0 + ...$, y en forma similar los pesos de orden superior. Por simplicidad, restringiremos nuestro análisis a los primeros ordenes en la expansión (4.2). Si las funciones a representar es suave, esta aproximación es suficiente para alcanzar buenos resultados, de lo contrario es posible ensayar algún desarrollo alternativo a (4.2), series de Fourier, por ejemplo. La incorporación de los pesos de orden superior le confiere a la red mayor poder de computo y es además susceptible a la aplicación de método de entrenamiento de Máxima Entropía con una leve modificación.

La idea ahora es introducir el algoritmo desarrollado en el capítulo previo [9] para seleccionar los pesos adecuados para captar la regla de un dado conjunto de entrenamiento. Primero construimos un vector **S** cuyas componentes son $\{1, \xi_i\}$,

los términos de segundo orden $\{\xi_i^2, \xi_i\xi_j\}$, y los de tercer orden $\{\xi_i^3, \xi_i^2\xi_j, \xi_i\xi_j\xi_k\}$. El conjunto de P patrones de entrada ξ^μ, con $\mu = 1, ..., P$ tiene ahora asociada la matriz $\mathbf{S}$ de esta forma el patrón de entrada captura las correlaciones del sistema en forma natural. El umbral de activación y los pesos w_i, w_{ij}, y w_{ijk} pueden ser escritos en un único vector $\mathbf{W}$ el cual satisface

$$g^{-1}(\zeta_0^\mu) = \mathbf{S}^\mu \cdot \mathbf{W}. \tag{4.3}$$

La ecuación (4.3) constituye nuestra información disponible.

La configuración de pesos más probable, compatible con los ejemplos (4.3), está dada en términos de la pseudoinversa de la matriz de $\mathbf{S}^\mu$

$$\mathbf{W} = I_{ps}[\mathbf{S}^\mu] g^{-1}(\zeta_0^\mu). \tag{4.4}$$

Este resultado puede ser inmediatamente generalizado a redes con más neuronas en la capa de salida. Para un mapeo dado por $\zeta_j = g(\mathbf{S} \cdot \mathbf{W}_j)$, la prescripción de Máxima Entropía es dada por

$$\mathbf{W}_j = I_{ps}[\mathbf{S}^\mu] g^{-1}(\zeta_j^\mu). \tag{4.5}$$

Esta será la receta a usar en los problemas siguientes.

4.4 Reconstrucción de funciones de onda

Nuestro objetivo en esta sección es aplicar la metodología de redes neuronales a aprender a construir con bastante aproximación la función de onda del estado fundamental a partir del conocimiento de solo algunos valores de expectación. Para ello empleamos el Principio de Máxima Entropía [10, 11, 12, 13, 14] de Jaynes, en dos formas muy diferentes: para optimizar la hipótesis de trabajo en el proceso de aprendizaje, y por otro lado, en construir en forma aproximada la función de onda del estado fundamental.

En la presente instancia, explicaremos brevemente el espíritu del último de estos ingredientes. La posibilidad de emplear solo un conjunto reducido de valores de expectación relevantes para determinar el operador estadístico de un sistema físico, constituye la razón de ser de la Mecánica Estadística. Sin embargo, si uno desea

aplicar un tratamiento similar para describir un estado puro, nos encontramos con una seria dificultad. Para estos estados la entropía de von Neumann-Shannon es identicamente nula. La forma de inferir estadisticamente la función de onda ha sido estudiado recientemente en diversos contextos [15, 16, 17, 18, 19, 20, 21, 22, 23, 24].

4.4.1 Entropía cuántica

Consideremos por simplicidad el problema unidimensional y construyamos la entropía "cuántica" S_Q introducida en [15, 16], la cual está definida por

$$S_Q = -\int |\psi|^2 \ln |\psi|^2 \, d\tau. \tag{4.6}$$

La distribución de probabilidades está asociada con el módulo de la pertinente función de onda. La entropía S_Q depende de la base y no es invariante con respecto a una transformación unitaria que cambie la base. La base a ser utilizada esta determinada por la naturaleza del los valores de expectación dados como fuente de información

$$\langle f \rangle = -\int |\psi|^2 f \quad d\tau. \tag{4.7}$$

Asumimos que los valores de expectación se refieren a operadores conmutantes, por lo tanto usamos la base en la cual ellos son diagonales. Maximización de S_Q, sujeta a las ligaduras (4.7) nos conduce al familiar problema de extremalización y finalmente a una forma exponencial para la función de onda

$$\psi(x) = \exp\left[-\frac{1}{2}\left(\lambda^0 + \sum_{l=1}^{L} \lambda^l x^l\right)\right], \tag{4.8}$$

donde los multiplicadores de Lagrange son obtenidos resolviendo el conjunto de ecuaciones acopladas

$$\frac{\partial \lambda^0}{\partial \lambda^l} = -\langle x^l \rangle . \tag{4.9}$$

Para el estado fundamental la función de onda no tiene nodos y (4.8) provee la aproximación de Máxima Entropía a la función de onda [15]. Para estados exitado, se requieren algunas consideraciones más elaboradas [23], que no son de interés aquí.

La Teoría de la Información provee la forma funcional de función de onda con un conjunto de multiplicadores de Lagrange a ser determinados resolviendo un complicado de ecuaciones diferenciales acopladas. Las redes neuronales constituye una alternativa válida para la determinación de los parámetros que evita el tedioso proceso de resolución de (4.9) para la determinación de los multiplicadores de Lagrange.

4.4.2 Aplicación específica

Como un primer ejemplo de aplicación, consideremos dos casos diferentes de un problema unidimencional, el potencial anarmónico ($V(x) = \alpha x^2 + \beta x^3 + \gamma x^4$) y el potencial de Morse ($V(x) = A(1 - \exp(-x))^2$). El correspondiente Hamiltoniano puede ser escrito como

$$\widehat{H} = \frac{\widehat{P}^2}{2} + V\left(\widehat{X}\right). \tag{4.10}$$

La prescripción para la función de onda de Máxima Entropía es [15]

$$\psi(x) = \exp\left[-\frac{1}{2}\left[\lambda^0 + \sum_{l=1}^{L} \lambda^l x^l\right]\right], \tag{4.11}$$

y los valores de expectación disponibles están dados por $\{< x^l >, l = 1, ..., L\}$. Los parámetros λ^l serán calculados por nuestra red debidamente entrenada, mientras que λ^0 es una constante de normalización. Funciones de onda de buena calidad son obtenida con solo tomar $L = 4$.

La performance de nuestro algoritmo ha sido estudiada en el caso de una red con función de transferencia lineal $g(x) = x$ y cuyo potencial de membrana h incorpora acoplamiento de segundo y tercer orden. El conjunto de entrenamiento es dado por pares *entrada–salida*. La salida consiste en P vectores $\overrightarrow{\lambda}_\mu = \{\lambda^l_\mu,\ l = 1, ..., 4\}$, con $\mu = 1, ..., P$. Dentro de un dado intervalo (ver Tabla 4.III), elegimos al azar con probabilidad uniforme los λ^l_μ y asociamos a cada vector $\overrightarrow{\lambda}$ la función de onda ψ_μ dada por (4.11). Por otro lado, las entradas son preparadas de la siguiente manera. Primero calculamos los momentos del estado fundamentas de la función de onda ψ_μ asociada con el conjunto $\overrightarrow{\lambda}_\mu$

$$< x^l_\mu >= \int \psi^*_\mu(x)\, x^l\, \psi_\mu(x)\, dx \qquad l = 1, .., 4. \tag{4.12}$$

	λ^1	λ^2	λ^3	λ^4
λ_{min}	-1.500	-0.700	-0.500	0.300
λ_{max}	1.500	0.700	0.350	1.100

Tabla 4.III: Intervalo de los valores de λ^l empleados para entrenar la red neuronal.

A primer orden la matriz de entrada **S** está dada por $\{1, < x^l >\}$; a segundo orden, $\{1, < x^l >, < x^l >< x^m >\}$, mientras que a tercer orden tenemos la matriz de entrada es dada por $\{1, < x^l >, < x^l >< x^m >, < x^l >< x^m >< x^n >\}$.

Primero estudiaremos la performance de las redes que incorporan diferentes ordenes en los acoplamientos. El error de generalización E_g está definido en término de alguna medida de la desviación ϵ entre la salida deseada λ^l_{exact} y la salida de la red λ^l correspondiente a una dada entrada $\langle x^l \rangle$. Como la sensibilidad de la función de onda a los parámetros λ_l no es la misma, la contribución al error debe ir pesada apropiadamente. Definimos como medida de la desviación a la cantidad

$$\epsilon = \frac{1}{2} \sum_{l=1}^{4} \left[\left| \langle x^l \rangle \right| \left(\lambda^l - \lambda^l_{exact} \right)^2 \right] . \tag{4.13}$$

El error de generalización E_g es el promedio de la desviación (4.13), el cual es calculado sobre 30 nuevos ejemplos (no pertenecientes al conjunto de entrenamiento). En la Fig. 4.2 podemos apreciar como a medida que la red incorpora acoplamientos de orden superior, el E_g disminuye significativamente (en un factor de 10) lo cual se traduce en una mejora sustancial en la performance de la red. La performance de la red disminuye hasta un valor crítico (meseta) diferente en cada tipo de red. Encontramos, además, que la cantidad de ejemplos necesarios para que el E_g alcance la meseta (saturación) es mayor cuando el orden de las interacciones aumenta. Sin embargo, si el número de ejemplos presentados a la red es debidamente escaleado con el número de neuronas de entrada N, obtenemos que con una cantidad de ejemplos del orden de N, la red esta en condiciones de generalizar con una excelente performance.

En las Fig. 4.3. mostramos las funciones de onda inferidas por una red con acoplamientos de tercer orden y la función de onda exacta, en dos casos diferentes. El acuerdo alcanzado es excelente.

Como un ejemplo más complicado, focalizaremos ahora nuestra atención en la

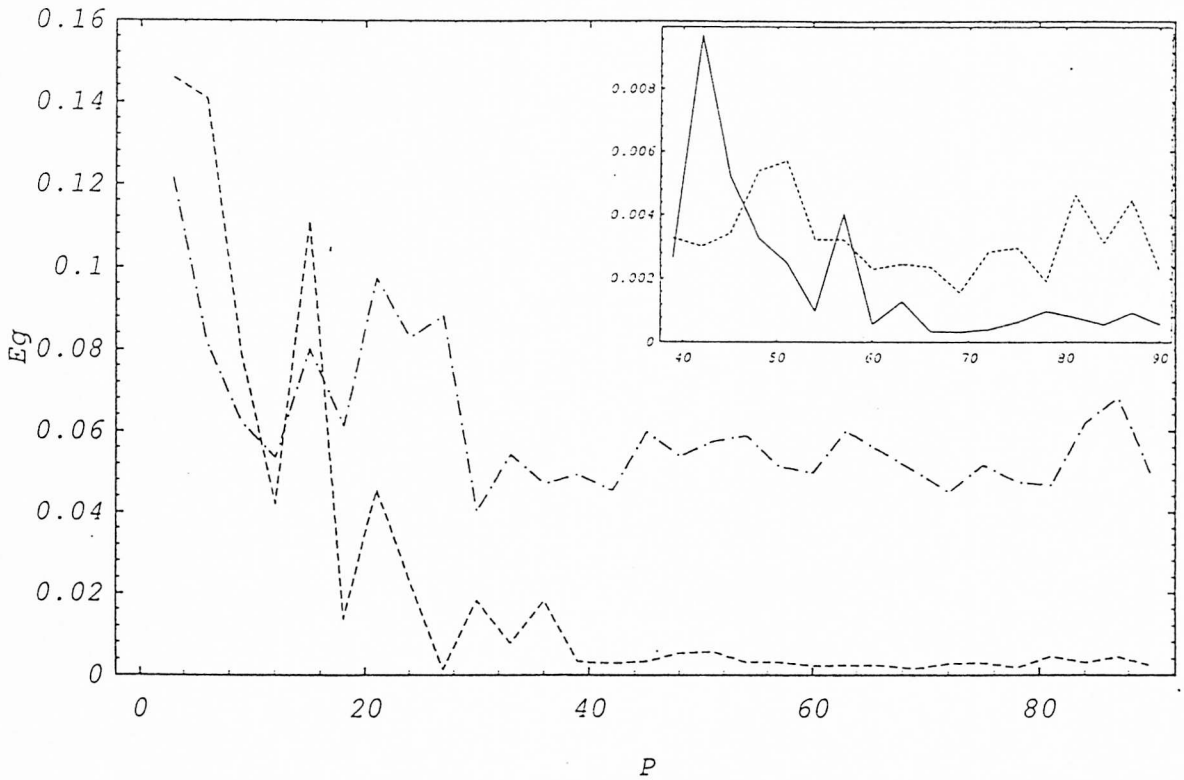

Figura 4.2: Error de generalización calculado sobre 30 nuevos ejemplos, como una función del número de ejemplos P usados en el entrenamiento para tres tipos de redes (ellas difieren en el orden de sus acoplamiento). En línea punto–raya repersenta los resultado de un simple perceptrón. La línea a rayas es una red que incorpora acoplamientos de segundo orden. A la dereche podemos comparar el error de generalización con acoplamientos de tercer orden, línea llena, con la línea a raya de acoplamiento de segundo orden.

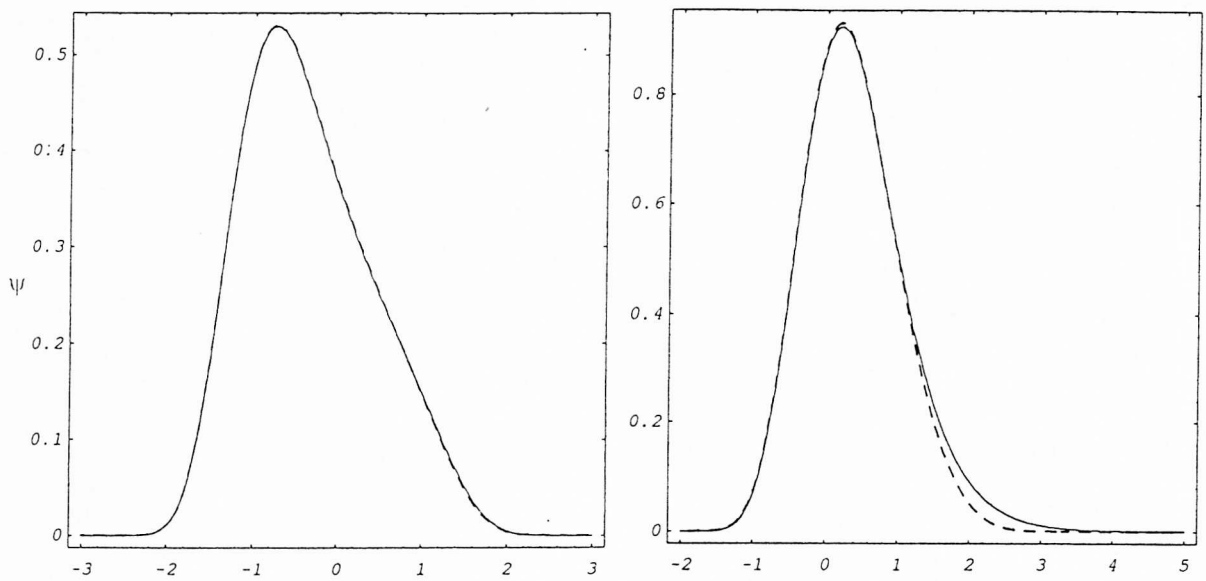

Figura 4.3: Izquierda: Función de onda del estado fundamental para un oscilador anarmónico, la inferida en línea a rayas y la exacta con línea llena. A la derecha: Podemos ver las funciones de onda para el potencial de Morse. La red fue entrenada con 63 ejemplos.

ecuación radial de Schrödinger, en el potencial de Coulomb y también en uno más complicado

$$V(r) = a_2 r^2 + a_3 r^3 + a_5 r^5 + a_8 r^8 \tag{4.14}$$

con $a_2, a_3, a_5, a_8 (a_8 > 0)$ arbitrarios. Para un potencial radial en tres dimensiones $V(r)$, escribimos la ecuación radial

$$\left[\frac{d^2}{dr^2} + U(r) + E\right] R(r) = 0, \tag{4.15}$$

con $U(r) = V(r) + \frac{l(l+1)}{2r^2}$. La parte radial de la función de onda de Máxima Entropía es $(\psi_r(r) = R(r)/r)$

$$\psi_r(r) = r^l \exp\left[-\frac{1}{2}\left[\lambda^0 + \sum_{k=1}^{L} \lambda^k r^k\right]\right]. \tag{4.16}$$

En la Fig. 4.4 mostramos las funciones de onda del estado fundamental para cada uno de los casos, junto con la solución exacta. El acuerdo obtenido es excelente.

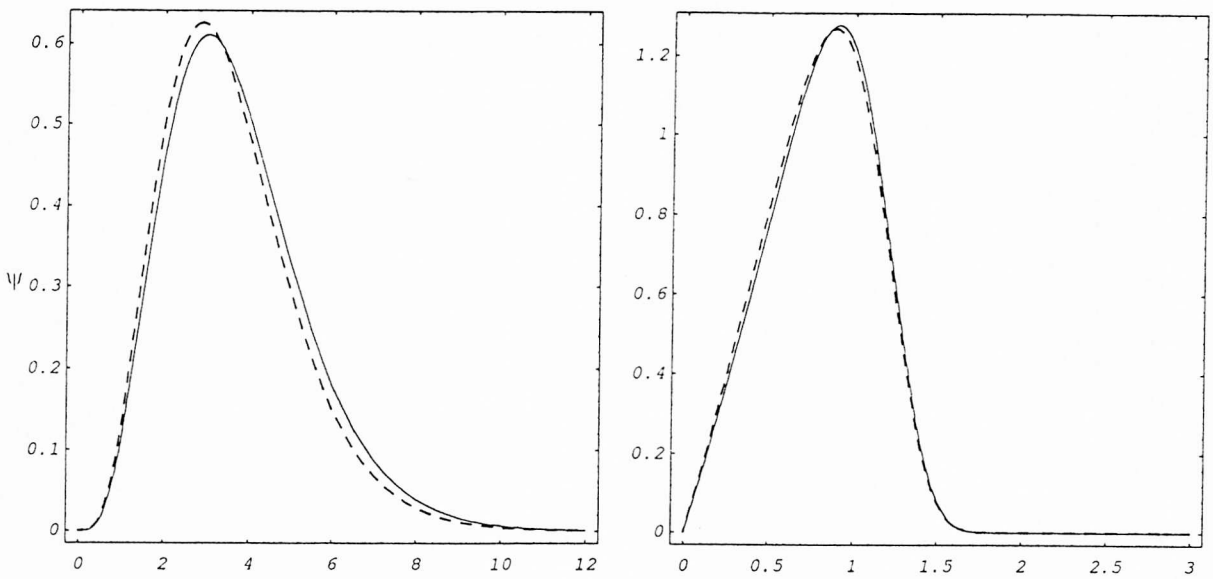

Figura 4.4: Izquierda: Función de onda del estado fundamental de un potencial de Coulomb, la inferida en línea a rayas y la exacta con línea llena. A la derecha: Podemos ver las funciones de onda para el potencial (4.14). La red fue entrenada con 63 ejemplos.

4.5 Predicción de series temporales caóticas

Uno de los objetivos centrales de la ciencia es la predicción. Dadas ciertas condiciones o eventos en el pasado, como predecir el futuro? La forma clásica de proceder consiste en construir un modelo explicatorio a partir leyes fundamentales y datos experimentales. Lamentablemente, esto no es siempre posible, ya sea porque no disponemos de un modelo en base a los Principios Fundamentales o porque la obtención de los datos experimentales precisos es muy dificultosa. Un alternativa para la predicción es la construcción de un modelo *ad hoc*, a partir de una secuencia temporal de datos [25]. Por esta razón mucho de los trabajos en análisis de la dinámica de sistemas no lineales han concentrado su atención en como construir un modelo apropiado a partir de una serie temporal caótica para predecir el comportamiento del sistema en un futuro próximo [26, 27, 28]. Ejemplos de sistema caóticos aparecen en muchos campos de física: fluidos turbulentos, reacciones química, laseres, física del plasma, ecosistemas, etc. En general, la mayoría de los modelos que permiten predecir son en base funciones nolineales parametrizadas [25], tales como modelo local lineal [26],

aproximación mediante funciones radiales [29, 30]. También muchos autores han aplicado redes neuronales a este problema [7, 31]. El objetivo en la presente sección es emplear la metodología previamente desarrollada en redes neuronales a la predicción de series temporales caóticas. Los sistemas caóticos muestran un comportamiento complejo y constituyen un excelente test para las habilidades predictivas de una red neuronal. Si uno puede enseñarle a la red el algoritmo determinista de una serie temporal caótica, entonces uno puede ser capaz de predecir el futuro de la serie temporal con bastante precisión.

4.5.1 Construcción del espacio de fase

Asumiremos que la señal es provista por un sistema dinámico $D : \Re^s \rightarrow \Re^s$ y la serie temporal consiste en una secuencia estroboscópica de medidas $\{v(t_n)\}$ del sistema en el estado $\mathbf{x}(t_n) \in \Re^s$ a tiempos discretos t_n, con $n = 1, \ldots, N$, siendo N la longitud de la serie. Si la dinámica es de baja dimensión podemos reconstruir el espacio de estados usando un *embedding* [32]. Whitney [33] mostró que un n-dimensional manifold M^n puede ser embebido en $\Re^k$ si $k \geq 2n$. Podemos reconstruir el espacio de estados por medio de las coordenadas retardadas, definimos $\Phi : \Re^s \rightarrow \Re^d$,

$$\mathbf{v}(t) = \Phi[\mathbf{x}(t)] \tag{4.17}$$

$$\mathbf{v}(t) = (v(t), v(t+\Delta), \ldots, v(t+(d-1)\Delta)), \tag{4.18}$$

donde Δ es el tiempo de *retardo* y d es la dimensión del *embedding* de esta reconstrucción. Takens [34] ha mostrado genéricamente que $d \geq 2s+1$, entonces Φ es un difeomorpismo entre los estados de $\Re^s$ y el espacio de coordenadas retardadas $\Re^d$. Uno puede obtener la dinámica de $v(t)$ a partir de la dinámica de los estados $\mathbf{x}(t)$ y del difeomorfismo $\mathbf{v}(t) = \Phi[\mathbf{x}(t)]$,

$$\mathbf{v}(t+T) = \Phi[\mathbf{x}(t+T)] = \Phi \circ D^T(\mathbf{x}(t)) \tag{4.19}$$

$$= \Phi \circ D^T \circ \Phi^{-1}(\mathbf{v}(t)) = \mathbf{F}(\mathbf{v}(t)). \tag{4.20}$$

La proyección de $\mathbf{v}(t+T)$ sobre una coordenada nos da $v(t+T) = \pi_1[\mathbf{F}(\mathbf{v}(t))]$. Obviamente, debido a la naturaleza caótica del sistema la exactitud de los pronósticos $v(t+T)$ decrece con T.

El teorema de Takens provee una dimensión entera suficiente d para el *embedding*, pero no es siempre necesaria. Para el atractor de Lorenz el teorema sugiere $d = 5$ (puesto que $d_a = 2.06$). Sin embargo, el método de los falsos vecinos cercanos [35] nos dice que la menor dimensión que puede desplegar el atractor sin ambigüedad es $d = 3$. Como el tiempo de computo de la matriz de pseudoinversa crece exponencialmete con d, elegimos la menor dimensión de *embedding* suficiente para desplegar completamente el atractor, la cual sera determinada por el método de los falsos vecinos cercanos a partir de los datos mismos .

El *embedding* de coordenadas de retardo (4.18) no es el único método para la reconstrucción de la dinámica. En muchos casos, la serie proviene de medidas experimentales contaminadas con ruido y las coordenadas de retardo no suele ser un buen *embedding*. Por esta razón, exploraremos también un sistema de coordenadas alternativo el cual considera variables filtradas [36]. El *embedding* de coordenadas filtradas consiste en la variable misma, en un filtro "sumador" y en un filtro "diferencial", es decir, la variable $\mathbf{v}(t_n) = (v_1(t_n), v_2(t_n), v_3(t_n))$ está dada por

$$\begin{aligned} v_1(t_n) &= v(t_n), \\ v_2(t_n) &= \sum_{j=n-2}^{n} v(t_j), \\ v_3(t_n) &= v(t_{n-1}) - v(t_n) \end{aligned} \tag{4.21}$$

debemos destacar que estos procesos de filtrado son introducidos para disminuir los efectos del ruido. El método de los vecinos falsos es robusto y determina correctamente la dimensión si el ruido presente en la señal no es superior al 5%. En esta sección, usaremos las coordenadas retardadas (4.18) y las variables filtradas (4.21) para la reconstrucción del espacio de estados a partir de la información de las medidas escalares. El último *embedding* solo sera usado para señales con ruido.

Por otro lado, el teorema de Takens no provee información acerca del retardo Δ a ser usado en el *embedding*. El tiempo de retardo puede ser en principio elegido casi arbitrariamente, pero si Δ es muy grande, entonces $v(t_0 + j\Delta)$ y $v(t_0 + (j+1)\Delta)$ son independientes una de otra, pero si el Δ usado en la reconstrucción es muy pequeño, entonces $v(t_0 + j\Delta)$ y $v(t_0 + (j+1)\Delta)$ serán indistinguibles en presencia

de ruido. Existe un criterio para la elección de [37], el cual establece que el retardo satifactorio es el primer mínimo de la información mutua [38, 39, 40]. La información mutua I_M está definida por

$$I_M(T) = \sum_n P(v(n), v(n+T)) \ln \left[\frac{P(v(n), v(n+T))}{P(v(n)) P(v(n+T))} \right] \tag{4.22}$$

donde $P(v(n))$ es la probabilidad de encontrar un dado valor v en la serie temporal, y $P(v(n), v(n+T))$ es la probabilidad condicional. Estudiaremos las bondades de nuestros pronósticos para diferentes tiempos de retardo.

4.5.2 Aplicación específica

Nuestro objetivo es ejemplificar la metodología de redes neuronales en la predicción de serie caóticas, para ello usaremos el modelo de Lorenz [41] y el mapa de Hénon [43]. Los datos generados numéricamente, son divididos en dos partes. Una de ellas es empleada para el entrenamiento de la red mientras que la otra es usada para testear el poder predictivo de la red. La tarea consiste en tomar un conjunto de datos

$$\{v_1(t_i), v_2(t_i), ..., v_d(t_i); v(t_i+T)\} \tag{4.23}$$

donde $i = 1, ..., P$ siendo P el número de datos de la serie temporal usados para entrenar a la red neuronal, la cual tendrá salida $v(t+T)$, y las entradas estarán dadas por $v_1(t), v_2(t), .., v_d(t)$.

Sistema de Lorenz

Uno de los sistemas nolineales más estudiado, proviene de la truncación de un sistema de ecuaciones diferenciales que describen la convección térmica en la baja atmósfera. Las ecuaciones resultantes son conocidas como sistema de Lorentz

$$\begin{aligned} \dot{x} &= c(y-x), \\ \dot{y} &= x(1-z)-y, \\ \dot{z} &= xy-bz, \end{aligned} \tag{4.24}$$

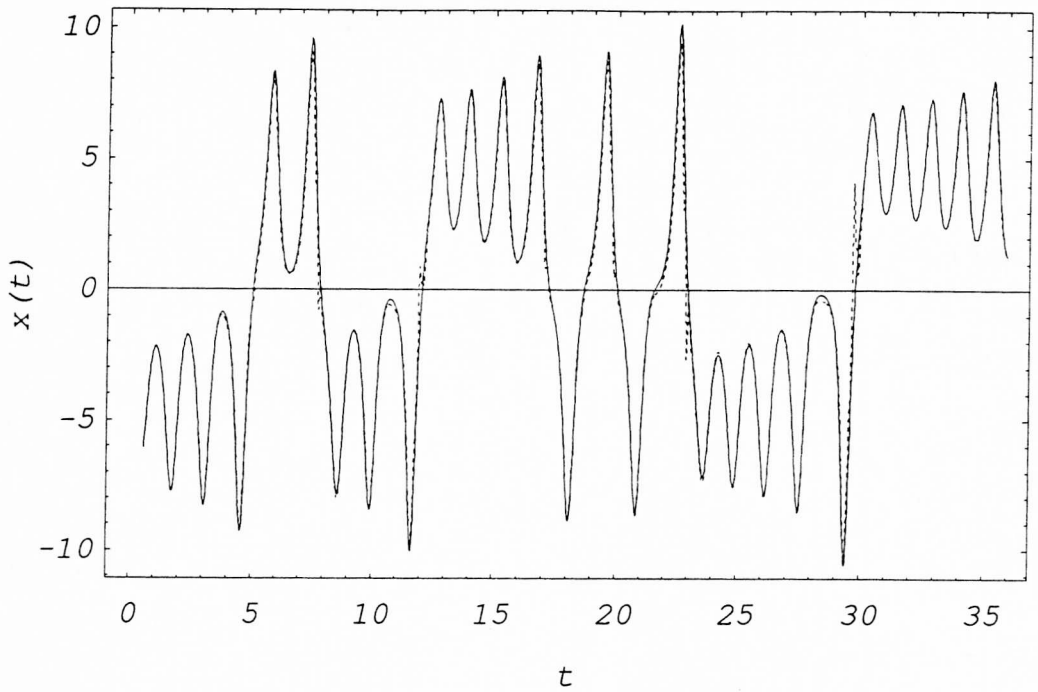

Figura 4.5: Solución númerica (línea llena) y predicciones (línea de trazos) para la variable x del sistema de Lorenz, con $T = 0.20$ y $\Delta = 0.1$. Solo 90 puntos han sido usados en el entrenamiento.

donde c es la razón entre la disipación térmica y la viscosa y b es un número sin dimensión que depende de la geometría de la celda convectiva. En el presente caso, hemos tomado $c = 3.5$ y $b = 2/3$. Para obtener los datos necesarios tanto para el modelado como para el testeo, resolvemos numéricamente las ecuaciones (4.24), usando el método de Runge-Kutta, con un paso 0.02. La dimensión del *embedding* usada en este problema es 3 [42], y la red neuronal incorpora acoplamiento de segundo y tercer orden. Solo 90 puntos de la serie son usados en el entrenamiento (es decir $M = 90$).

En la Fig. 4.5, podemos ver la solución numérica $x(t)$ de la ecuación (4.24) como así también nuestra predicciones con $T = 0.20$ y $\Delta = 0.1$. Por supuesto la bondad del pronóstico, como en todo sistema caótico, decrece con el aumento del T. En la Fig. 4.6 mostramos el error cuadrático medio como una función de T para diferentes retardos. El error crece en forma exponencial con T.

También analizamos la performance del método para diferentes retardos Δ. En la Fig. 4.7, podemos apreciar el error cuadrático medio como una función del retardo.

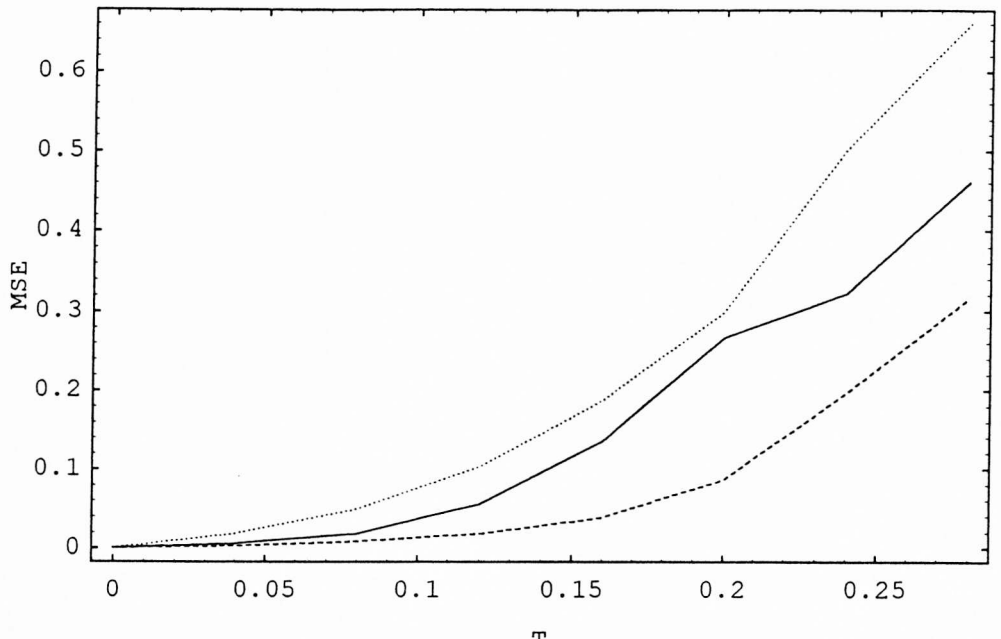

Figura 4.6: Error cuadrático medio vs. el tiempo de predicción T.

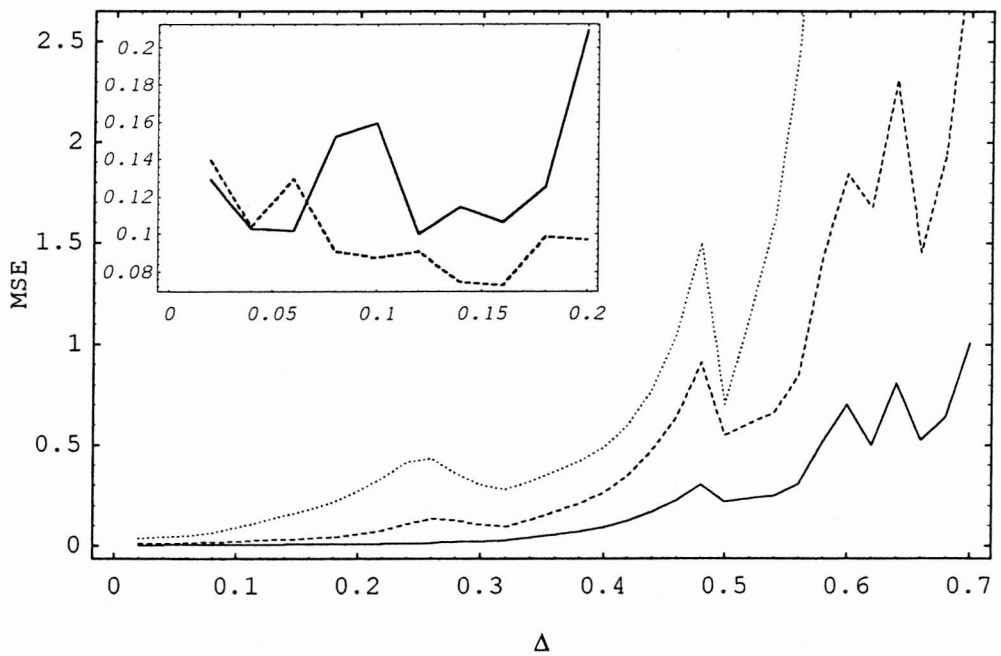

Figura 4.7: Error cuadrático medio como función del tiempo de retardo Δ. La línea llena corresponde a $T = 0.04$, la línea de trazo a $T = 0.12$ y la línea de puntos a $T = 0.2$. En el recuadro, la línea de trazo corresponde a al *embedding* filtrado y la llena al retardado. En este caso $T = 0.04$.

Los pronósticos son bastante buenos, cuando el retardo es pequeño. Para ilustrar el procedimiento con señales ruidosas, usamos otra vez el atractor de Lorenz con ruido Gaussiano adicional, de media cero y varianza 0.5. Para este nivel de ruido, el porcentaje de vecinos falsos es menor de $\approx 2\%$, y elegimos 3 como dimensión del *embedding*. En la Fig. 4.7, podemos comparar el error cuadrático medio de los dos *embedding* (4.18) y (4.21), (ver recuadro). Evidentemente, en este caso el *embedding* de variables filtradas es mejor que las variables de coordenadas retardadas. Es evidente que cuando la señal esta contaminada con ruido un Δ pequeño no es una buena elección. Como podemos apreciar, una buena elección es alrededor de $\Delta = 0.1$, en buen acuerdo con el criterio de Fraser.

Mapa de Hénon

Consideremos aquí la serie temporal x_i, generada por el mapa de Hénon

$$\begin{aligned} x_{n+1} &= 1 + y_n - a\, x_n^2, \\ y_{n+1} &= b\, x_n \end{aligned} \tag{4.25}$$

donde $a = 2$ y $b = 0.3$. La dimensión del *embedding* dimensión usado es $d = 2$, mientras que el teorema de Takens asegura que el atractor se despliega con $d = 3$, puesto que la dimensión del atractor para este mapa es 1.26. En la Fig. 4.8 podemos ver la variable x_n, tanto la provista por (4.25) como las predicciones de la red neuronal. En este caso acuerdo alcanzado es excelente.

4.6 Conclusiones

En este capítulo, hemos investigado dos ejemplos de la aplicación directa de una red neuronal con unidades continuas, entrenada con el algoritmo de Máxima Entropía. Esta red es capaz de aprender correctamente una regla a partir de un pequeño conjunto de ejemplos. Nuestra técnica garantiza un aprendizaje perfecto de los ejemplos (error de entrenamiento cero) y una excelente performance en la generalización. La simple minimización de la energía de entrenamiento sólo garantiza una buena performance para los patrones miembros del conjunto de entrenamiento, pero no nos

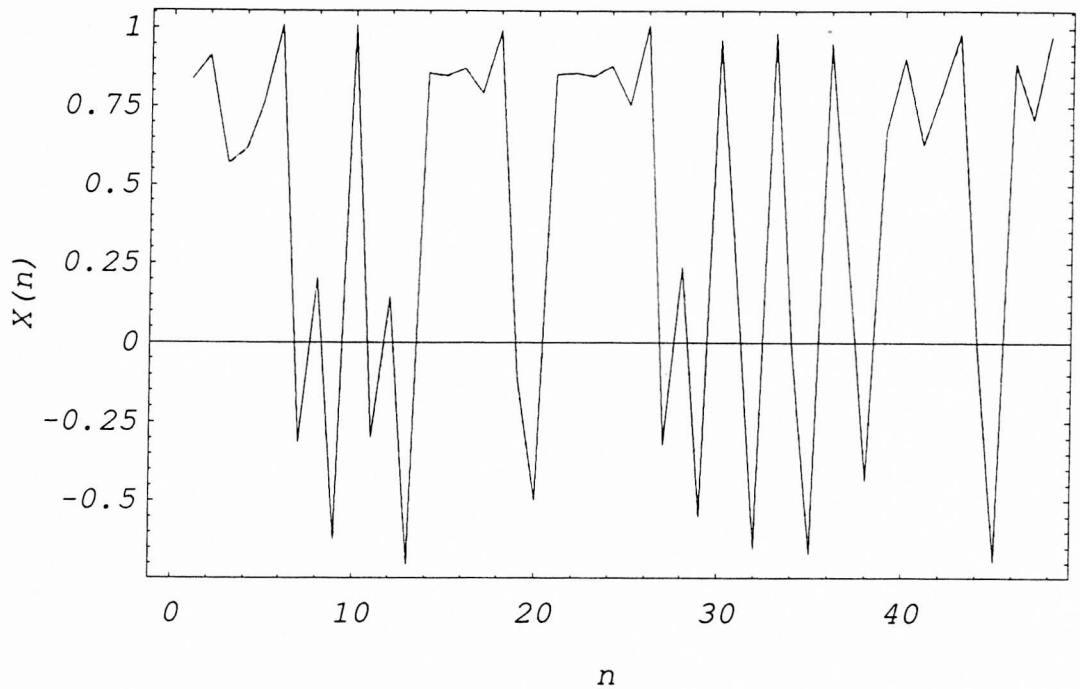

Figura 4.8: Pronostico de la red neuronal ($T = 2$ iteraciones) (línea de trazo) y la serie (línea sólida) del mapa Hénon.

asegura nada acerca de la generalización. Un hecho remarcable en el esquema aquí presentado es la pequeña cantidad de ejemplos necesarios para lograr una excelente performance en el proceso de entrenamiento. Esta es una característica notable que diferencia el esquema aquí presentado, de los métodos más ortodoxos, donde además la convergencia del proceso de aprendizaje puede ser muy lenta y se necesitan una enorme cantidad de ejemplos para lograr una generalización aceptable. Nuestro formalismo puede ser fácilmente aplicado a una variedad de problemas.

Hemos ilustrados nuestras consideraciones con referencia a dos problemas importantes de la física: la inferencia de funciones de onda y las predicción de series temporales. En el primer caso, la red aprende la relación entre el conjunto de valores de expectación y los multiplicadores de Lagrange asociados a una función de onda, sin ningún conocimiento en lo concerniente al potencial de interacción. Este esquema puede ser implementado sin muchas modificaciones para la reconstrucción de estados excitados y del potencial de interacción asociado. Por otro lado, en lo concerniente al modelado y predicción de series temporales caóticas nuetros resultados presentan una excelente perspectiva. No es novedosa la aplicación de redes neuronales a la

predicción, pero si es de interés remarcar la pequeña cantidad de ejemplos necesaria para hacer pronósticos de calidad. Mientras que en otros esquemas (retropropagación de error, por ejemplos) son requeridos del orden 1000 o más ejemplos [44], aquí sólo hemos usado apenas un centenar.

Resumiendo, a lo largo de los últimos dos capítulos, se ha dejado claro que la Teoría de la Información enmarcada en el Principio de Máxima Entropía, ofrece una nueva alternativa para el desarrollo de algoritmos de entrenamiento más efectivos para redes neuronales.

Bibliografía

[1] F. Rosemblatt, *Principles of Neurodynamics* (Spartan, New York 1962).

[2] D.E. Rumelhart y J.L. McClelland, *Parallel Distributed Processing* (MIT, Cambridge, MA 1988).

[3] H.S. Seung, H. Sompolinsky y N. Tishby, Phys. Rev. A **45**, 6056 (1982).

[4] M. Opper y D. Haussler, Phys. Rev. Lett. **66** , 2677 (1991).

[5] T. Watkin, A. Rau y M. Biehl, Rev. Mod. Phys. **65**, 499 (1993).

[6] L. Diambra y A. Plastino, Phys. Rev. E **52**, 4557 (1995).

[7] L. Diambra y A. Plastino, Phys. Rev. E **53**, 1021 (1996).

[8] I.J. Matus y P. Perez, Phys. Rev. A **43**, 5683 (1991).

[9] L. Diambra, J. Fernandez y A. Plastino, Phys. Rev. E **52**, 2887 (1995).

[10] C.E. Shannon y W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Chicago, Ill 1949).

[11] E.T. Jaynes, Phys. Rev. **106**, 620 (1957); **108**, 171 (1957).

[12] R.D. Levine y M. Tribus, *The Maximum Entropy Principle* (MIT Press, Boston MA 1978).

[13] A. Katz, *Principles of Statistical Mechanics* (Freeman, San Fransisco 1967).

[14] N. Agmon, Y. Alhassid y R.D. Levine, en *The Maximum Entropy Formalism* editado por R.D. Levine y M. Tribus (MIT Press, Cambridge 1979).

[15] N. Canosa, A. Plastino y R. Rossignoli, Phys. Rev. A **40**, 519 (1989).

[16] A.R. Plastino y A. Plastino, Phys. Lett. A **181**, 446 (1993).

[17] N. Canosa, R. Rossignoli y L. Diambra, Phys. Lett. A **185**, 133 (1994).

[18] N. Canosa, A. Plastino y R. Rossignoli, Nucl. Phys. A **512**, 550 (1990).

[19] N. Canosa, A. Plastino, R. Rossignoli y H.G. Miller, Phys. Rev. C **45**, 1162 (1992).

[20] L. Arrachea, N. Canosa, A. Plastino, M. Portesi y R. Rossignoli, Phys. Rev. A **45**, 7104 (1992).

[21] N. Canosa, A. Plastino y R. Rossignoli, Nucl. Phys. A **550**, 453 (1992).

[22] L. Arrachea, N. Canosa, A. Plastino, y R. Rossignoli, Phys. Lett. A **176**, 353 (1993).

[23] M. Casas, A. Plastino, y A. Puente, Phys. Rev. A **49**, 2312 (1994).

[24] M. Casas, A. Plastino y A. Puente, Phys. Lett. A **184**, 385 (1994).

[25] H.D.I Abarbanel, R. Brown, J.J. Sidorowich, y L.S. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).

[26] J.D. Farmer y J.J. Sidorowich, Phys. Rev. Lett. **59**, 845 (1987).

[27] M. Casdagli, Physica D **45**, 335 (1989).

[28] M. Giona, F. Lentini, y V. Cimagalli, Phys. Rev. A **44**, 3496 (1991).

[29] J. Moody y C. Darken, Neural Comput. **1**, 281 (1989).

[30] X. He, y A. Lapedes, Physica D **70**, 289 (1993).

[31] A. Lapedes y R. Farber, *Neural Information Processing Systems*, editado por D.Z. Anderson (AIP, New York) p.442 (1987).

[32] T. Sauer, J.A. Yorke, y M. Casdagli, J. Stat. Phys. **65**, 579 (1991).

[33] H. Whytney, *Geometric Integration Theory*, (Princeton University Press, Princeton, New Jersey, 1957).

[34] F. Takens, *Lecture notes in Mathematics*, Vol 898 (Springer, Berlin), 366 (1981).

[35] M.B. Kennel, R. Brown, y H.D.I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).

[36] G.B. Mindlin, H.G. Solari, M.A. Natiello, R. Gilmore, y X.-J. Hou, J. Nonlinear Sci. **1**, 147 (1991).

[37] A.M. Fraser, IEEE Trans. Inf. Theory **35**, 245 (1989).

[38] K. Ikeda y K. Matsumoto, Phys. Rev. Lett. **62**, 2265 (1989).

[39] K. Kaneko, Physica D **41**, 137 (1990).

[40] J.-L. Chern, Phys. Rev. E **50**, 4315 (1994).

[41] E.N. Lorenz, J. Atmos. Sci. **20**, 130 (1963); **26**, 636 (1969).

[42] L. Diambra y A. Plastino, aceptado para su publicación en Phys. Lett. A.

[43] M. Hénon, Commun. Math. Phys. **50**, 69 (1976).

[44] A. Lapedes y R. Farber, LA-UR-87-2662 (1987).

Apéndice A

Cálculo de la función de generalización

La función de generalización está definida según (2.3), por

$$\varepsilon(\mathbf{W}) = \frac{1}{2}\int d\mu(\mathbf{S})\left[g(\mathbf{W}_0\cdot\mathbf{S}) - g(\mathbf{W}\cdot\mathbf{S})\right]^2, \tag{A.1}$$

introduciendo las variables auxiliares x y y, removemos $\mathbf{S}$ de los argumentos de las funciones g

$$\int dxdy\frac{1}{2}\left[g(x) - g(y)\right]^2 \int d\mu(\mathbf{S})\delta\left(x - N^{-1/2}\mathbf{W}\cdot\mathbf{S}\right)\delta\left(x - N^{-1/2}\mathbf{W}_0\cdot\mathbf{S}\right) \tag{A.2}$$

donde $d\mu(\mathbf{S})$ denota una medida sobre el espacio de las entradas. Para poder seguir es necesario establecer la medida de integración. A lo largo esta tesis, solo consideraremos el caso en que las entradas están distribuidas independientemente en la forma $d\mu(\mathbf{S}) = \prod_i^N \frac{dS_i}{2\pi}\exp\left(-\frac{S_i^2}{2}\right)$, reemplazando en (A.2), y usando la identidad $\delta(x) = (2\pi)^{-1}\int dx' e^{ixx'}$ obtenemos

$$\begin{aligned}&\int\frac{dxdy}{2\pi}\int\frac{dx'dy'}{2\pi}\left[g(x) - g(y)\right]^2 e^{ixx'+iyy'}\times\\ &\int\frac{d\mathbf{S}}{2\pi}\exp\left[\frac{\mathbf{S}^2}{2} - iN^{-1/2}\left(\mathbf{W}x' + \mathbf{W}_0y'\right)\cdot\mathbf{S}\right],\end{aligned} \tag{A.3}$$

completando cuadrados, integramos fácilmente la integral sobre los ejemplos

$$\exp\left[-\left(\frac{1}{2}x'x' + y'y' + x'y'R\right)\right], \tag{A.4}$$

donde $R = N^{-1}\mathbf{W}_0\cdot\mathbf{W}$ es el solape entre los vectores pesos del PM y PE. La integral sobre las variables primadas da finalmente

$$\varepsilon(R) = \frac{1}{4\pi}\int\frac{dxdy}{\sqrt{1-R^2}}\exp\left[\frac{x^2+y^2-2xyR}{2(1-R^2)}\right]\left[g(x) - g(y)\right]^2. \tag{A.5}$$

Como podemos ver, el comportamiento de la función de generalización está determinado solamente por el parámetro de orden R.

Cálculo de la correlación

En el cálculo de la correlación (2.18) interviene solamente $C1_{\gamma\delta}$ el cual esta dado por

$$C1_{\gamma\delta} = \frac{1}{4}\int d\mu\left(\mathbf{S}\right)\left[g\left(\mathbf{W}_{\gamma}\cdot\mathbf{S}\right)-g\left(\mathbf{W}_{0}\cdot\mathbf{S}\right)\right]^{2}\left[g\left(\mathbf{W}_{\delta}\cdot\mathbf{S}\right)-g\left(\mathbf{W}_{0}\cdot\mathbf{S}\right)\right]^{2}, \tag{A.6}$$

en forma similar eliminamos las variables $\mathbf{S}$ del argumento de la función de transferencia,

$$\begin{aligned}&\int\ dxdydz\frac{1}{4}\left[g\left(x\right)-g\left(z\right)\right]^{2}\left[g\left(y\right)-g\left(z\right)\right]^{2}\times \\ &\int\ d\mu\left(\mathbf{S}\right)\delta\left(x-N^{-1/2}\mathbf{W}_{\gamma}\cdot\mathbf{S}\right)\delta\left(y-N^{-1/2}\mathbf{W}_{\delta}\cdot\mathbf{S}\right)\delta\left(z-N^{-1/2}\mathbf{W}_{0}\cdot\mathbf{S}\right).\end{aligned} \tag{A.7}$$

Haciendo $\delta\left(x\right)=\frac{1}{2\pi}\int dx'\exp\left(ixx'\right)$ en (A.7)

$$\begin{aligned}&\frac{1}{4}\int\int\frac{d\mathbf{r}d\mathbf{r}'}{\left(2\pi\right)^{3}}\exp\left(i\mathbf{r}\cdot\mathbf{r}'\right)\left[g\left(x\right)-g\left(z\right)\right]^{2}\left[g\left(y\right)-g\left(z\right)\right]^{2}\times \\ &\int d\mu\left(\mathbf{S}\right)\exp\left[-iN^{-1/2}\left(\mathbf{W}_{\gamma}x'+\mathbf{W}_{\delta}y'+\mathbf{W}_{0}z'\right)\cdot\mathbf{S}\right].\end{aligned} \tag{A.8}$$

Reemplazando $d\mu\left(\mathbf{S}\right)$ y completando cuadrados en el último factor de (A.8), tenemos

$$\exp\left[-\left(\frac{1}{2}\mathbf{r}'\cdot\mathbf{r}'+x'z'R_{\gamma}+y'z'R_{\delta}+x'y'Q_{\gamma\delta}\right)\right], \tag{A.9}$$

donde

$$\begin{aligned}R_{\gamma}&=N^{-1}\mathbf{W}_{\gamma}\cdot\mathbf{W}_{0}\\ R_{\delta}&=N^{-1}\mathbf{W}_{\delta}\cdot\mathbf{W}_{0}\\ Q_{\gamma\delta}&=N^{-1}\mathbf{W}_{\gamma}\cdot\mathbf{W}_{\delta}.\end{aligned} \tag{A.10}$$

Completando cuadrado, podemos integrar fácilmente en las variables primadas,

$$\frac{1}{4\pi}\int\frac{dxdydz}{\left(2\pi\Xi_{\gamma\delta}\right)^{1/2}}\exp\left[-\frac{K\left(x,y,z\right)}{2\Xi_{\gamma\delta}}\right]\left[g\left(x\right)-g\left(z\right)\right]^{2}\left[g\left(y\right)-g\left(z\right)\right]^{2} \tag{A.11}$$

donde

$$\begin{aligned}K\left(x,y,z\right)&=2xy\left(R_{\gamma}R_{\delta}-Q_{\gamma\delta}\right)-2zx\left(R_{\gamma}-R_{\delta}Q_{\gamma\delta}\right)-2zy\left(R_{\delta}-R_{\gamma}Q_{\gamma\delta}\right)\\ &\quad z^{2}\left(1-Q_{\gamma\delta}^{2}\right)+y^{2}\left(1-R_{\gamma}^{2}\right)+x^{2}\left(1-R_{\delta}^{2}\right)\\ \Xi_{\gamma\delta}&=\left(1-R_{\gamma}^{2}\right)\left(1-R_{\delta}^{2}\right)-\left(R_{\gamma}R_{\delta}-Q_{\gamma\delta}\right)^{2}\end{aligned}$$

Apéndice B

Cálculos relativos al perceptrón Booleano

Para el perceptrón de salida booleana $g(x) = \text{sign}(x)$ definiremos la función error como $\epsilon(\mathbf{W},\mathbf{S}) = \theta\left(-N^{-1/2}(\mathbf{W}\cdot\mathbf{S})(\mathbf{W}_0\cdot\mathbf{S})\right)$. En este caso, la función de generalización puede ser escrita, después de un cambio de variable, como

$$\int_0^\infty \int_{-\infty}^0 \frac{dxdy}{\pi\sqrt{1-R^2}} \exp\left[-\frac{x^2+y^2+2xyR}{2(1-R^2)}\right], \tag{B.1}$$

pasando a coordenadas polares obtenemos

$$\frac{1}{\pi}\int_{\pi/2}^{\pi}\int_0^\infty \frac{d\theta\rho d\rho}{\sqrt{1-R^2}} \exp\left[-\frac{\rho^2(1+2R\sin\theta\cos\theta)}{2(1-R^2)}\right], \tag{B.2}$$

integrando sobre ρ, obtenemos finalmente

$$\varepsilon(R) = \frac{1}{\pi}\int_\pi^{2\pi} \frac{d\theta}{\sqrt{1+R\sin\theta}} \equiv \frac{1}{\pi}\cos^{-1}(R). \tag{B.3}$$

La definición usual de la función error como la desviación cuadrática, este resultado difiere solo en un factor 2. El cálculo de $C1_{\gamma\delta}$ tiene mayor labor algebraica, pero afortunadamente tiene una expreción analítica cerrada, en termino de los parámetros R y $Q_{\gamma\delta}$. Para el caso booleano, tomamos nuevamente la función error definida en el parrafo anterior, en este caso tenemos

$$C1_{\gamma\delta} = 2\int_0^\infty\int_0^\infty \frac{dxdy}{2\pi}\exp\left[-\frac{y^2(1-R_\gamma^2)+x^2(1-R_\delta^2)+2xy(R_\gamma R_\delta - Q_{\gamma\delta})}{2\Xi_{\gamma\delta}}\right]\times$$
$$\int_{-\infty}^0 \frac{dz}{(2\pi\Xi_{\gamma\delta})^{1/2}}\exp\left[-\frac{z^21-Q_{\gamma\delta}^2-2z(x(R_\gamma-R_\delta Q_{\gamma\delta})+y(R_\delta-R_\gamma Q_{\gamma\delta}))}{2\Xi_{\gamma\delta}}\right]. \tag{B.4}$$

La integral en z da

$$\left[\frac{\pi\Xi_{\gamma\delta}}{2(1-Q_{\gamma\delta}^2)}\right]^{1/2}\text{erfc}\left[\frac{x(R_\gamma-R_\delta Q_{\gamma\delta})+y(R_\delta-R_\gamma Q_{\gamma\delta})}{(2\Xi_{\gamma\delta}(1-Q_{\gamma\delta}^2))^{1/2}}\right], \tag{B.5}$$

donde $\text{erfc}(x) = \frac{2}{\pi}\int_x^\infty e^{-t^2}dt$. Pasando a coordenadas polares podemos escribir (B.4) como

$$\int_0^{\pi/2}\int_0^\infty \frac{\rho d\rho d\theta}{2\pi(1-Q_{\gamma\delta}^2)^{1/2}}\exp\left[-\frac{a\rho^2(1-Q_{\gamma\delta}\sin 2\theta)}{2\Xi_{\gamma\delta}(1-Q_{\gamma\delta}^2)}\right]\text{erfc}(\rho c) \tag{B.6}$$

donde

$$\begin{aligned} a &= 1 - Q_{\gamma\delta}^2 - R_\gamma^2 - R_\delta^2 + 2R_\delta R_\gamma Q_{\gamma\delta} \\ c(\theta) &= \frac{(R_\gamma - R_\delta Q_{\gamma\delta})\cos\theta + (R_\delta - R_\gamma Q_{\gamma\delta})\sin\theta}{\left(2\Xi_{\gamma\delta}\left(1 - Q_{\gamma\delta}^2\right)\right)^{1/2}} \end{aligned} \tag{B.7}$$

integrando en ρ tenemos

$$\frac{\Xi_{\gamma\delta}\left(1 - Q_{\gamma\delta}^2\right)^{1/2}}{2\pi a}\int_0^{\pi/2} d\theta \left[\frac{1}{1 - Q_{\gamma\delta}\sin 2\theta} + \sqrt{\frac{c^2(\theta)}{c^2(\theta) + \frac{a\left(1 - Q_{\gamma\delta}\sin 2\theta\right)}{2\Xi_{\gamma\delta}\left(1 - Q_{\gamma\delta}^2\right)}}}\right]. \tag{B.8}$$

Resolviendo estas integrales, llegamos al resultado

$$C1_{\gamma\delta} = \frac{1}{2\pi}\left[\tan^{-1}\left[\frac{Q_{\gamma\delta}}{\left(1 - Q_{\gamma\delta}^2\right)^{1/2}}\right] + \pi - \sin^{-1}\left[1 - R_\gamma^2 - R_\delta^2\right]\right]. \tag{B.9}$$

Si suponemos simetría de réplica $R_\gamma = R_\delta = R$ y $Q_{\gamma\delta} = \begin{cases} 1 & \gamma = \delta \\ q & \gamma \neq \delta \end{cases}$. Hay dos tipos de aportes a considerar en (B.9) los n términos diagonales

$$\frac{n}{2\pi}\left(\frac{3\pi}{2} - \sin^{-1}\left[1 - 2R^2\right]\right), \tag{B.10}$$

mientras que los $n^2 - n$ términos fuera de la diagonal suman (despreciando los terminos de orden n^2)

$$-\frac{n}{2\pi}\left(\tan^{-1}\left[\frac{q}{(1 - q^2)^{1/2}}\right] + \pi - \sin^{-1}\left[1 - 2R^2\right]\right). \tag{B.11}$$

Como podemos ver en el límite $n \to 0$ los términos $\sin^{-1}\left[1 - 2R^2\right]$, no aportan a la energía de correlación lo cual justifica la ecuación (2.32) y escribir

$$\sum_{\gamma\delta}^{n} C1_{\gamma\delta} = \frac{n}{2\pi}\left(\frac{\pi}{2} - \tan^{-1}\left[\frac{q}{(1 - q^2)^{1/2}}\right]\right). \tag{B.12}$$

Apéndice C

Correlación cuando las salidas son azar

Consideraremos ahora el cálculo las correlaciones (2.56), cuando las salidas t están dadas al azar. La función error, en este caso es $\epsilon(\mathbf{W},\mathbf{S},t) = \theta\left(-N^{-1/2}(\mathbf{W}\cdot\mathbf{S})\,t\right)$. La energía de correlación estará dada por

$$C2_{\gamma\delta} = \int D(t)\,dt D(\mathbf{S})\,d\mathbf{S}\,\theta\left(-N^{-1/2}(\mathbf{W}_\gamma\cdot\mathbf{S})\,t\right)\,\theta\left(-N^{-1/2}(\mathbf{W}_\delta\cdot\mathbf{S})\,t\right), \qquad \text{(C.1)}$$

por medio del artificio de la delta de Dirac, removemos las variables $\mathbf{S}$

$$\int\frac{dxdx'}{2\pi}\int\frac{dydy'}{2\pi} \qquad \int D(t)\,dt\,\theta(-tx)\,\theta(-yt)\exp\left[i\left(xx'+yy'\right)\right]\times$$
$$\int\frac{d\mathbf{S}}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\mathbf{S}^2+2i\mathbf{S}\cdot\left(\mathbf{W}_\gamma x'+\mathbf{W}_\delta y'\right)N^{-1/2}\right)\right] \qquad \text{(C.2)}$$

donde hemos usando la representación $\delta(x) = (2\pi)^{-1}\int dx'\exp(ixx')$ de la función de Dirac y tomamos una medida gaussiana $D(\mathbf{S}) = (2\pi)^{-1}\exp(-\mathbf{S}^2/2)$, la integración sobre $d\mathbf{S}$ conduce al resultado (intermedio)

$$\exp\left[-\left(\frac{1}{2}(x')^2+(y')^2+2x'y'Q_{\gamma\delta}\right)\right], \qquad \text{(C.3)}$$

donde $Q_{\gamma\delta} = N^{-1}\mathbf{W}_\gamma\cdot\mathbf{W}_\delta$. El cálculo sobre las variables primadas da

$$\left(\frac{1-Q_{\gamma\delta}^2}{2\pi}\right)^{1/2}\int\frac{dxdydt}{2\pi}\theta(-xt)\,\theta(-yt)\exp\left[-\frac{1}{2}\left(t^2+x^2+y^2-2xyQ_{\gamma\delta}\right)\right] \qquad \text{(C.4)}$$

integrando en t, llegamos a

$$\left(1-Q_{\gamma\delta}^2\right)^{1/2}\int_0^\infty\frac{dxdy}{2\pi}\exp\left[-\frac{1}{2}\left(x^2+y^2-2xyQ_{\gamma\delta}\right)\right]$$

pasando a coordenadas polares, e integrando obtenemos

$$C2_{\gamma\delta} = \frac{1}{4}+\frac{1}{2\pi}\tan^{-1}\left[\frac{Q_{\gamma\delta}}{\left(1-Q_{\gamma\delta}^2\right)^{1/2}}\right]. \qquad \text{(C.5)}$$

Tomando en cuanta, la simetría de réplica, encontramos que los n términos diagonales, suman $n/4$, mientras que para los restantes $n^2 - n$ términos, tenemos (despreciando los términos de orden $O(n^2)$),

$$= \frac{n}{2\pi}\left(\frac{\pi}{2} + \tan^{-1}\left[\frac{q}{(1-q^2)^{1/2}}\right]\right), \tag{C.6}$$

por lo tanto

$$\sum_{\gamma\delta}^{n} C2_{\gamma\delta} = \frac{n}{2\pi}\left(\frac{\pi}{2} - \tan^{-1}\left[\frac{q}{(1-q^2)^{1/2}}\right]\right). \tag{C.7}$$

en igual forma que a $C1_{\gamma\delta}$ (B.12). A segundo orden en β, las contribuciones son idénticas, tanto para los ejemplos erroneos, como para los provistos por el PM. Cada contribución esta pesada por la cantidad de ejemplos de cada tipo y provienen de la aleatoriedad de los ejemplos, la cual esta presente en los dos tipos de ejemplos.

Esta tesis está basada en las siguientes publicaciones:

- Capítulo 2 :

  Thermodynamics of nonlinear perceptron,
  L. Diambra, M.T. Martín, C. Mostaccio, y A. Plastino,
  preprint: La Plata-Th 96/2.

  Perturbative treatment and learning Techniques,
  L. Diambra y A. Plastino, en prensa.

- Capítulo 3 :

  Pseudoinverse techniques, information theory, and the training of the feedforward networks,
  L. Diambra, J. Fernández, y A. Plastino,
  Phys. Rev. E **52**, 2887 (1995).

  Neural network training without spurious minima,
  L. Diambra y A. Plastino, en prensa.

- Capítulo 4 :

  The Maximum-entropy principle and neural networks that learn to construct approximate wave functions,
  L. Diambra y A. Plastino,
  Phys. Rev. E **53**, 1021 (1996).

  Maximum entropy, pseudoinverse techniques and time series predictions with layered networks,
  L. Diambra y A. Plastino,
  Phys. Rev. E **52**, 4557 (1995).

  Time series modelling using information theory,
  L. Diambra y A. Plastino, en prensa.