Facultad de Ciencias

Departamento de Materia Condensada

TRABAJO FIN DE GRADO

Inteligencia Artificial aplicada a transiciones de fase.

Hugo Pérez Martínez

Director: Luis Martín Moreno

Universidad Zaragoza

Curso 2019/2020

Índice

1. Abstract

En el presente trabajo utilizamos redes neuronales para llevar a cabo la predicción de la transición de fase del modelo de Ising bidimensional. Utilizamos dos tipos de estructuras para las redes neuronales: redes *fully connected* y convolucionales. La obtención de la temperatura crítica que caracteriza la transición se realiza de dos maneras diferentes. Por un lado entrenamos a las redes en la predicción del orden de las configuraciones de Ising, situando la transición en la temperatura cuyas configuraciones pasan de considerarse mayoritariamente ordenadas a desordenadas. Por otro lado, centramos el entrenamiento en la predicción de la longitud de correlación, estableciendo la temperatura crítica como aquella cuyas configuraciones dan lugar a las mayores predicciones. Los resultados obtenidos son en general satisfactorios; muestran que las redes *fully connected* centran sus predicciones en propiedades globales de la red como la magnetización, aunque desconocen la topología local de la misma dando lugar a malos resultados en la longitud de correlación, mientras que las convolucionales se comportan de manera complementaria, obviando los observables globales y sobreestimando el valor de la transición con el método de clasificación al carecer de información sobre la magnetización.

Nota del autor. El campo de la inteligencia artificial es relativamente joven y se ha desarrollado principalmente en la literatura anglosajona. Como consecuencia, no existen buenas traducciones al castellano para la mayoría de los términos utilizados en el campo, y se suelen utilizar con su formulación original. Por tanto, de ahora en adelante recurriremos a los nombres originales en inglés en la mayoría de ocasiones, salvo que la traducción sea evidente e inequívoca. Por otro lado, por decisión de estilo, el presente trabajo se ha escrito utilizando la primera persona del plural pese a tener un único autor.

2. Introducción

El concepto de Inteligencia Artificial (IA) engloba todas aquellas actividades llevadas a cabo por una máquina que reproducen comportamientos que consideramos inteligentes por parte de humanos o animales. Se trata por tanto de un campo extremadamente amplio que abarca una gran cantidad de actividades, desde procesamiento de lenguaje hasta conducción autónoma, pasando por el reconocimiento de imágenes. Para llevar a cabo todos estos procedimientos es necesario que las inteligencias artificiales se someta a un proceso de aprendizaje, dando lugar a una nueva rama de investigación, el Aprendizaje Automático o *Machine Learning* (ML), en la que se han desarrollado multitud de herramientas y algoritmos que logran muy buenos resultados para aplicaciones cada vez más complejas. El desarrollo actual de la Inteligencia Artificial se basa en la disponibilidad y accesibilidad de una cantidad cada vez mayor de datos e información, parte crucial en el proceso de aprendizaje. La Inteligencia Artificial está pues llamada a convertirse en una de las principales herramientas tecnológicas del futuro.

La Inteligencia Artificial no sólo es aplicable en la vida cotidiana, también comienza a aparecer en la ciencia. En los últimos años ha crecido exponencialmente la disponibilidad de datos

experimentales y la capacidad de procesamiento, y la posibilidad de clasificar y extraer información de forma rápida y precisa resulta cada vez más atractiva. En este sentido se están planteando IAs capaces por ejemplo de clasificar la ingente cantidad eventos provenientes de colisiones de partículas en el LHC [1], predecir estructuras proteicas a partir de la secuencia primaria del DNA, o reproducir correctamente el comportamiento de fluidos sin recurrir a costosas simulaciones numéricas. En el campo de la materia condensada, y concretamente en el estudio de diversos modelos estadísticos como el modelo de Ising, la Inteligencia Artificial nos permite predecir una gran cantidad de propiedades y clasificar configuraciones según diversos criterios.

El objetivo principal de este trabajo consistirá pues en construir redes neuronales capaces de predecir la existencia de una transición de fase en el modelo de Ising. Utilizaremos para ello dos propiedades diferentes. Por un lado, crearemos redes capaces de discernir entre configuraciones ordenadas y desordenadas, según si corresponden a temperaturas menores o mayores que la crítica. Por otro, construiremos una red para predecir la longitud de correlación de cualquier configuración, que diverge en la transición. En ambos casos, generando grandes cantidades de configuraciones para muchas temperaturas y realizando las predicciones, podemos encontrar el valor crítico en el que se cumplen las condiciones para la transición, a saber: el punto en el que las configuraciones pasan de clasificarse mayormente como ordenadas a desordenadas y viceversa, y el punto en el que las configuraciones tienen las mayores longitudes de correlación.

La estructura del trabajo es la siguiente: en la primera parte se discuten los conceptos básicos que utilizaremos durante el trabajo. En la sección 3 presentamos las redes neuronales, las estrategias principales que se utilizan durante el entrenamiento y los posibles problemas que pueden surgir. En la sección 4 realizamos un breve resumen de las propiedades principales del modelo de Ising que se utilizarán en el trabajo. En la segunda parte se presentan los resultados obtenidos; en la sección 5 se muestran los resultados para las redes neuronales que predicen la transición de fase mediante clasificación de configuraciones, y en el apartado 6 se muestran los resultados para las redes neuronales que predicen la longitud de correlación. Por último, en el apartado 7 se discuten las conclusiones más importantes del trabajo.

Parte I
Conceptos básicos

3. Redes neuronales

Una de las estrategias más populares en la aplicación del ML son las llamadas "redes neuronales", inspiradas en las estructuras que forman las neuronas en los seres vivos, que han dado lugar a toda una nueva rama de estudio sobre *Machine Learning* con algoritmos específicos de aprendizaje. La unidad fundamental de este tipo de redes es la neurona, que se coloca en la red formando parte de una capa. Las redes constan de una o varias capas sucesivas entre la información de entrada o *input*, y la de salida o *output*. En los últimos años han ganado protagonismo las redes con multitud de capas intermedias, dando lugar a lo que conocemos como *deep*

learning. Las capas intermedias reciben el nombre de *hidden layers* o *deep layers*. Describimos a continuación los conceptos básicos que utilizaremos durante el trabajo.

3.1. Neuronas

Las neuronas son pequeñas unidades de la red capaces de recibir un *input* de la capa anterior y generar un *output* para la capa siguiente, aplicando función $f(z)$ sobre dicho *input*. Cada neurona k se caracteriza a su vez por un conjunto de pesos $\mathbf{w_k^r} = (w_{k1}^r, \ldots, w_{kN}^r)$ correspondientes a cada una de sus uniones con las N neuronas de la capa anterior, y un *bias* b_k^r. El superíndice $r \in 1, ..., R$ indica la capa a la que corresponden dichas variables. De esta forma, dado un output $\mathbf{x^{r-1}} = (x_1^{r-1}, \ldots, x_N^{r-1})$ proveniente de la capa anterior, el valor que se introduce en la función propia de la neurona n viene dado por:

$$z_k^r = \sum_j w_{kj}^r x_j^{r-1} + b_k^r = \mathbf{w_k^r} \cdot \mathbf{x^{r-1}} + b_k^r \tag{1}$$

A partir de este input generamos un nuevo output $f(z_k^r) \equiv a_k^r$. Al atravesar toda la red obtenemos finalmente una predicción $\mathbf{a^R}(w, b)$[1], que es función de todos los pesos w y *biases* b de todas las capas. Omitimos los índices para indicarlo. Tenemos por tanto una gran cantidad de parámetros en la red que podemos ajustar: un *bias* por neurona y un peso por cada unión entre neuronas. Esto hace que incluso en redes relativamente simples tengamos que ajustar miles de parámetros, y en las más complejas hasta millones. Aquí reside la potencia de estas redes, que les permite adaptarse a una gran cantidad de problemas diferentes y reproducir resultados complejos, pero también es uno de sus principales inconvenientes. Primero, es muy difícil y costoso computacionalmente trabajar con una cantidad tan elevada de parámetros, y los tiempos de entrenamiento pueden resultar prohibitivos. Segundo, corremos el riesgo de sobreajustar el comportamiento de la IA, de forma que más allá de aprender las características generales del problema para predecir correctamente sobre nuevos datos, comience a aprender las características particulares del conjunto de datos utilizados como entrenamiento, dando lugar a predicciones erróneas o poca capacidad de generalización en un proceso conocido como *overfitting*. Por todo esto trataremos siempre de reducir al máximo el número de parámetros de nuestras redes. Veremos algunos métodos para lidiar con ambos problemas en apartados posteriores.

Exiten una gran cantidad de funciones posibles que podemos utilizar en una red neuronal. En nuestro caso aplicaremos las funciones *sigmoid*, *ReLU*, *leaky ReLU* y *softmax*. La primera es la pieza fundamental de gran parte del trabajo, y como su propio nombre indica, utiliza la sigmoide como función de activación:

$$f(z) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

La representación gráfica se muestra en la figura (1). El *output* puede tomar cualquier valor entre 0 y 1, siendo la función suave y derivable en todo su dominio. La forma de esta función otorga gran versatilidad al funcionamiento de la red siempre y cuando los *biases* y pesos den lugar a valores de z en torno al origen, donde se produce un mayor cambio en el output bajo pequeñas variaciones de input. Sin embargo, el hecho de que las derivadas en los extremos de la función

[1]Puede tratarse de un único escalar si la última capa tiene una sola neurona, o de un vector si tiene varias.

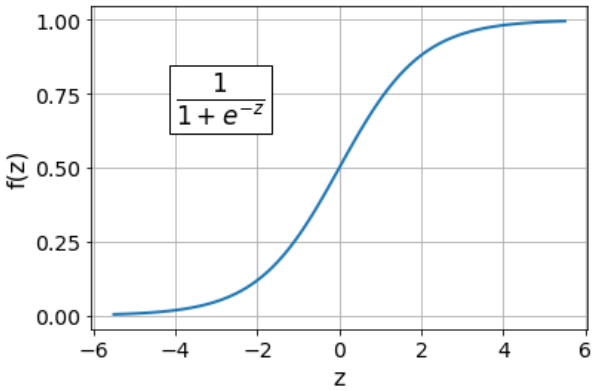


Figura 1: Función sigmoide.

(regiones planas) se anulen puede resultar muy problemático en la aplicación del algoritmo de *backpropagation* durante el proceso de aprendizaje (ver anexo 4).

Además de ser razonablemente versátiles y poder formar parte de casi cualquier red neuronal, el hecho de poseer una imagen entre 0 y 1 convierte a estas neuronas en óptimas para formar parte de la capa de salida en problemas de clasificación. En estas aplicaciones nos interesa obtener un *output* que podamos interpretar como "pertenece a una clase" o "no pertenece a una clase". Matemáticamente la aproximación más simple se da con 0 y 1, y basta construir una capa a la salida formada por tantas neuronas sigmoides como clases busquemos diferenciar. Aquella neurona con mayor *output* marca la predicción de la red. En caso de querer realizar predicciones en las que el *output* $\mathbf{y}$ no esté necesariamente contenido entre 0 y 1, necesitamos normalizar los datos de entrenamiento para que ninguno supere la unidad. Esto obliga a determinar de antemano un valor máximo para el *output*, asignarle la unidad y normalizar el resto frente a él.

3.2. Tipos de redes

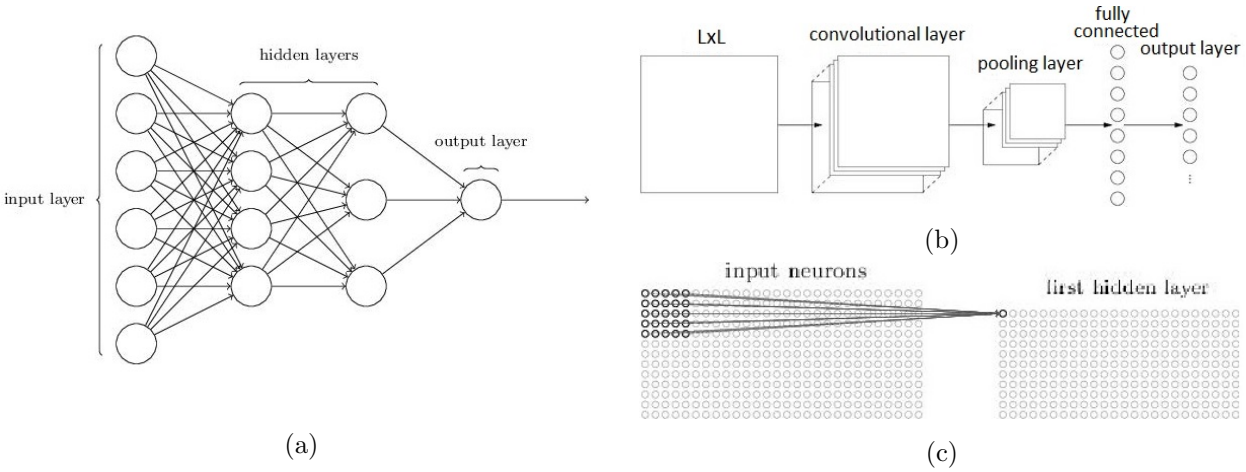
A lo largo del trabajo utilizaremos dos tipos de redes diferentes: redes completamente conectadas (*fully connected*), y redes convolucionales. En las primeras cada neurona recibe un input de todas las neuronas de la capa inmediatamente anterior, realiza una operación determinada, y emite un output a todas las neuronas de la capa inmediatamente posterior. Vemos un ejemplo de este tipo de redes en la figura (2a). Las redes convolucionales son más complejas. Utilizan capas intermedias con estructura de matriz formadas por neuronas cuyo *input* no procede de todas las neuronas de la capa anterior, sino únicamente de aquellas en su vecindad. El parámetro que determina el número de neuronas de la capa anterior utilizadas viene dado por el *receptive field*, de forma (l, l). Por ejemplo, si tenemos un *receptive field* $(5, 5)$, los *inputs* vendrán de la neurona anterior y sus dos primeros vecinos en todas las direcciones, incluyendo la diagonal. Vemos esto más claramente en la figura (2c). De esta forma, partiendo de una red de tamaño $L \cdot L$, la siguiente capa intermedia será otra red de, como máximo, $L \cdot L$ neuronas[2]. Tenemos

[2]En principio el tamaño de la red se reduce con cada capa convolucional debido al tamaño del *receptive field*. Por ejemplo, con un *receptive field* de $(5, 5)$ la siguiente capa debería tener un tamaño de $(L-2, L-2)$, debido a que no hay información más allá para llenar las dos capas restantes. Lo que se hace en ocasiones es rellenar la

entonces un número de pesos dado por el *receptive field* y un *bias*. La clave de este tipo de redes es que estos pesos y *biases* no son únicos para cada neurona, sino que todas ellas comparten los mismos valores. Siguiendo el ejemplo anterior con un *receptive field* de $(5, 5)$, tenemos un total de 25 pesos y 1 *bias*, y serán estos los que utilice cada neurona para determinar su *input*. Las ventajas evidentes de este tipo de redes es que reducimos en gran medida el número de parámetros necesarios, fomentamos la capacidad de generalización del resultado, y explotamos la posible invariancia traslacional de nuestros datos.

Es aconsejable además la utilización de las conocidas como *pooling layers*. Estas capas reducen el tamaño de la matriz con la que trabajamos agrupando la información de la capa anterior mediante diversos criterios, reduciendo aún más el número de parámetros. Funcionan de forma similar al resto de capas, pero en esta ocasión cada neurona de la capa anterior sólo puede servir de *input* para una neurona de la capa siguiente. Por ejemplo, si tomamos un *receptive field* de $(2, 2)$, el *output* de cada conjunto de cuatro neuronas formando un cuadrado únicamente servirá como *input* de una neurona. De esta manera reducimos el tamaño de la red a la mitad, $(L/2, L/2)$. Además, las neuronas de estas capas suelen tener un comportamiento muy simple; los casos más comunes son la *max pooling layer*, cuyo *output* es igual al mayor de sus *inputs*, y la *average pooling layer*, cuyo *output* es el valor promedio de sus *inputs*.

Las capas de la red pueden tener a su vez subcapas, denominadas filtros, que reciban la información de la misma capa anterior y apliquen pesos y *biases* diferentes. De esta manera, logramos capas tridimensionales de la forma (L, L, f), donde f es el número de filtros. Un mayor número de filtros permite la detección de más características mejorando el funcionamiento de la red. Además, estas capas se pueden superponer una tras otra, y cada filtro utilizará la información de las $l \cdot l \cdot f$ neuronas anteriores. Al finalizar es conveniente añadir una última capa *fully connected* antes de la capa de *output* para filtrar el resultado. Vemos una red que une todas las ideas anteriores en la figura (2b).



(a) (b) (c)

Figura 2: (a) Ejemplo de red neuronal *fully connected* con varias capas intermedias, capa de *input* y de *output*. (b) Ejemplo de red con una capa convolucional, una *pooling layer*, una capa *fully connected* y capa de *output*. (c) Concepto de campo receptivo en una red convolucional. En la imagen, las neuronas de la primera capa presentan un campo receptivo de (5,5). Imágenes editadas, fuente original: [2].

capa de *input* a los lados con ceros hasta poder crear una nueva capa de tamaño (L, L). Esto será lo que hagamos en el trabajo.

En el proceso de construcción de una red neuronal se deben fijar algunas características como el número de capas, el número de neuronas por capa, el tamaño del *receptive field*... Todas estas cantidades reciben el nombre de hiperparámetros (frente a los pesos y *biases*, que se denominan parámetros), se establecen desde el principio e influyen de forma clara en el funcionamiento de la red. Se requiere por tanto un estudio previo para determinar los valores que pueden tomar. En general el funcionamiento de las redes es muy sensible a los valores seleccionados, aunque en nuestro caso lograremos buenos resultados para casi cualquier combinación. Los comentados hasta ahora no son los únicos hiperparámetros, veremos más a lo largo del trabajo.

3.3. Entrenamiento

Nos fijamos primero en los datos que debe ser capaz de interpretar la IA. La información de entrada es un conjunto de $L \cdot L$ spines en el caso bidimensional, y $L \cdot L \cdot L$ en el tridimensional. Al trabajar con redes *fully connected*, los spines se introducen en la red como un único vector de datos de dimensión $(L \cdot L, 1)$ o $(L \cdot L \cdot L, 1)$ respectivamente. Para ello se toma la configuración de Ising y se construye un vector concatenando las filas una detrás de otra. De esta manera, las redes neuronales carecen en principio de información topológica acerca de la red de spines, salvo en la dirección horizontal. De ser relevante, esta información deberá adquirirla durante el proceso de aprendizaje. Por otro lado, en las redes convolucionales se introduce la red de Ising bidimensional completamente estructurada, de forma que la información topológica está presente desde el principio y la red neuronal se puede centrar en el comportamiento local de los spines.

Los principios que rigen el proceso de entrenamiento de una IA para lograr predicciones correctas son relativamente simples. Partimos de una red formada por neuronas conectadas, en las que los pesos y *bias* toman valores iniciales al azar. Tomamos una configuración y la introducimos en el sistema. Obtenemos una predicción a la salida que distará del resultado real que esperamos de la red, y variamos ligeramente los parámetros para lograr que el *output* se aproxime un poco más a dicho resultado. Repetimos este procedimiento varias veces para todas las configuraciones de entrenamiento hasta que obtenemos unas predicciones que consideramos correctas. El problema que se presenta en este proceso es el inmenso tamaño del espacio de fases de la red, que alcanza con facilidad las decenas o cientos de miles de parámetros dificultando el proceso de aprendizaje. Los largos periodos de tiempo necesarios para llevar a cabo el entrenamiento han sido históricamente el principal lastre para el desarrollo de IAs, aunque se han desarrollado multitud de estrategias que comentaremos a continuación.

3.3.1. Función coste

Para lograr el aprendizaje debemos definir en primer lugar una función que nos permita cuantificar el rendimiento de la red. Esta es la función coste $C(w, b)$, que nos da un valor numérico de la desviación del *output* $\mathbf{a}^{\mathbf{R}}(w, b)$ de la red respecto del valor esperado $\mathbf{y}$. Se calcula con el *output* real y el esperado, y por tanto es función del conjunto de pesos y *biases* de la red, lo que nos permite determinar con relativa facilidad cómo debemos variar estos para disminuir

el valor final sin más que determinar el gradiente del coste $\nabla C(w, b)$ frente a cada peso y *bias* individual. En problemas de clasificación la función coste más utilizada es la *cross-entropy*:

$$C = -\frac{1}{M} \sum_{x=1}^{M} [\mathbf{y} \cdot \ln \mathbf{a} + (1 - \mathbf{y}) \cdot \ln(1 - \mathbf{a})] \tag{3}$$

Las operaciones con términos vectoriales deben entenderse elemento a elemento, i.e. $\ln(1 - \mathbf{a}) \equiv (\ln(1 - a_1), \ldots, \ln(1 - a_N))$, siendo N el número de neuronas de la última capa. M es el número de configuraciones usadas en el entrenamiento, y la suma se da sobre todos sus *outputs*. Las propiedades principales de esta función y su conveniencia se discuten en el anexo 1. Independientemente de la función utilizada, el cálculo de las derivadas resulta muy complejo por la elevada cantidad de parámetros y la propia estructura de la red; basta darse cuenta de que un peso cualquiera en una capa intermedia influye en la función coste a través de todas las neuronas de las capas posteriores. Este problema encuentra fácil solución mediante la aplicación del algoritmo de *backpropagation*, que nos da una receta sencilla para el cálculo del gradiente recorriendo la red desde la última capa hasta la primera (de ahí su nombre), y para la actualización de los parámetros.

Otro problema reseñable es que al definirse la función coste sobre la totalidad de las configuraciones utilizadas en el entrenamiento, necesitamos calcular todos los *outputs* y compararlos con los reales antes de actualizar los parámetros. Esto puede resultar muy exigente desde el punto de vista computacional si trabajamos con una gran cantidad de configuraciones o redes relativamente grandes, ralentizando en gran medida el entrenamiento. La estrategia natural para combatir este problema consiste en asumir que el gradiente de la función coste calculado con la totalidad de las configuraciones es muy cercano al gradiente que podemos calcular haciendo uso de un subconjunto de las mismas. Naturalmente, cuanto mayor sea el subconjunto más se parecerán los gradientes, pero más lento será el proceso de entrenamiento. El tamaño de este subconjunto se conoce como *mini-batch*, y es un hiperparámetro de la red. Esta aproximación al problema se conoce como *stochastic gradient descent*, y se discute en el anexo 2.

3.3.2. *Overfitting*

La existencia de una función coste dependiente de los parámetros de la red nos permite también afrontar el problema del *overfitting* comentado anteriormente. Una gran cantidad de parámetros, lejos de dotar a la red de una mayor capacidad predictiva al aumentar su complejidad, puede llevar fácilmente a la identificación errónea de patrones, de forma que las predicciones no se basen en características generales del modelo de Ising sino en particularidades del conjunto de configuraciones de entrenamiento. Podemos entenderlo como un efecto análogo a los problemas que surgen al ajustar polinomios de diverso grado a un conjunto de datos experimentales. Con un mayor grado logramos más precisión en el ajuste, pero comprometemos la capacidad de predicción del resultado a obtener en otras medidas fuera del conjunto inicial de datos, y con toda probabilidad estaremos incluyendo el ruido de la medida. Trasladado al modelo de Ising, un indicador general que puede aportar información útil en la predicción de la fase es la magnetización, y es deseable que la red sea capaz de identificarla como tal. Sin embargo, si la

red basa sus predicciones en otras características como si la configuración comienza con un spin positivo o negativo, o si los grupos de spines orientados en la misma dirección tienden a ser más anchos que altos (ambas cosas podrían ocurrir en suficientes configuraciones entre los datos de entrenamiento como para que la red las identificara como factores), entonces las predicciones de la red se verían afectadas y perderían validez.

Existen varias aproximaciones posibles para combatir este problema. Por un lado, podemos aumentar el número de configuraciones que utilizamos para entrenar la red, de forma que las posibles peculiaridades de configuraciones concretas se vean diluidas y no afecten en gran medida al resultado final. Por otro lado, resulta conveniente reducir el número de parámetros para obligar a la red a generalizar su comportamiento. Podemos lograrlo reduciendo el tamaño de la red, pero esto puede ser muy perjudicial para su rendimiento, y es preferible recurrir a otra técnica llamada "regularización". Consiste en añadir un término adicional a la función coste original C_0 que penalice la existencia de una gran cantidad de parámetros y los posibles valores abultados que puedan tomar, $C(w,b) = C_0(w,b) + \lambda\ reg(w,b)$. El término de regularización viene pesado por el hiperparámetro de regularización λ; un valor elevado de λ reduce el número de parámetros útiles, y viceversa. Esto puede ir en detrimento del correcto funcionamiento de la red, y por tanto deberemos ser cuidadosos a la hora de seleccionar su valor. Este nuevo término de regularización puede tomar distintas funciones, y puede incluir o no los *biases*. En nuestro caso utilizaremos una función $L2 = \frac{1}{2M}\sum_w w^2$ y no incluiremos los *biases*. Discutimos estas elecciones en el anexo 1. La función coste resulta finalmente:

$$C = -\frac{1}{M}\sum_{x=1}^{M}[\mathbf{y}\cdot\ln\mathbf{a} + (1-\mathbf{y})\cdot\ln(1-\mathbf{a})] + \frac{\lambda}{2M}\sum_w w^2 \tag{4}$$

La última estrategia útil para evitar el *overfitting* consiste en detener el entrenamiento en el momento adecuado, lo que se denomina *early stopping*. Nos interesa que la función coste tome un valor final muy pequeño para asegurar que las predicciones son correctas, pero si exigimos un valor demasiado pequeño lo más probable es que la red comience a aprender las particularidades de los datos de entrenamiento para reducirlo, perjudicando su capacidad de generalización. Utilizaremos diferentes estrategias para abordar este problema.

4. Modelo de Ising

Hemos centrado el trabajo en la transición de fase que se da en el modelo de Ising en una red cuadrada a partir de dos dimensiones. Resulta ideal para este estudio por su simpleza y el amplio conocimiento disponible sobre su comportamiento para diferentes temperaturas. Describimos a continuación brevemente el modelo y sus características principales. El cálculo y los detalles de la simulación del modelo de Ising se discuten en el anexo 5.

El modelo de Ising consiste en un conjunto de spines $\{\sigma_\alpha\}$ que pueden tomar dos valores diferentes, $+1$ y -1, dando lugar a una configuración $C(\sigma_\alpha)$ caracterizada por una energía:

$$\mathcal{H}_\alpha = -\sum_{\langle i,j\rangle} J_{ij}\sigma_i\sigma_j - H\sum_i \sigma_i \tag{5}$$

En nuestro caso no aplicaremos campo magnético, y por tanto prescindimos del segundo término. La suma del primer término se realiza sobre todas las posibles parejas de spines $\langle i, j\rangle$ de la red, y viene pesada por un término J_{ij} que marca la intensidad de la interacción. En el trabajo nos ceñiremos al caso en el que la interacción se da únicamente entre primeros vecinos, y además es constante para todas las parejas. En caso de ser positiva, favorece energéticamente que los spines se alineen, y da lugar a un comportamiento ferromagnético. Si es negativa, favorece un comportamiento antiferromagnético. A lo largo del trabajo haremos uso de ambos. Esta energía fija la probabilidad de que aparezca una configuración determinada $C(\sigma_\alpha)$ a cierta temperatura T según la distribución de Boltzmann:

$$P(\{\sigma_\alpha\}, T) = \frac{1}{Z(T)} e^{-\frac{\mathcal{H}_\alpha}{k_B T}} = \frac{1}{Z(T)} e^{-\beta \mathcal{H}_\alpha} \tag{6}$$

De ahora en adelante utilizaremos la temperatura y $\beta = (k_B T)^{-1}$ indistintamente, donde k_B es la constante de Boltzmann, aunque tomaremos siempre $k_B = 1$. El término $Z(T)$ representa la función de partición del sistema a cierta temperatura, $Z(T) = \sum_{\{\sigma_\alpha\}} \exp(-\beta \mathcal{H}_\alpha)$, donde la suma se da sobre todas las posibles configuraciones del sistema. Definimos también la magnetización de una configuración como la suma de sus spines, $m = \frac{1}{N} \sum_j \sigma_j$, que es el parámetro de orden del modelo ferromagnético. En la fase desordenada, que se da en altas temperaturas, su valor esperado es 0. La naturaleza del modelo implica que los spines tienen dos orientaciones posibles que dan lugar a dos fases ordenadas, una en la que todos los spines toman valor $+1$, y otra en la que toman -1. Si bien ambas son simétricas en sus propiedades, son diferentes desde el punto de vista de la red neuronal, y deberemos asegurarnos de que en las configuraciones utilizadas para entrenamiento y predicción estén presentes ambos comportamientos. En el caso antiferromagnético es sencillo comprobar que el valor de este observable será prácticamente nulo para todas las temperaturas. En este caso utilizamos la *staggered magnetization*, esto es, la magnetización de las dos subredes que se forman tomando como vecinos los spines en diagonal en lugar de los adyacentes.

En ambos casos, la transición de fase en el límite termodinámico (esto es, para redes infinitas) se produce cuando el valor esperado de los observables respectivos deja de ser nulo, y se trata de una transición de segundo orden. Nosotros trabajaremos con redes finitas imponiendo condiciones de contorno periódicas para suplir en parte esta carencia. En el caso bidimensional y con $J = \pm 1$, Onsager [4] demostró que la transición de fase se da en $\beta_c^{2d} \simeq 0{,}44069$. Nuestro objetivo en el trabajo será lograr predecir esta temperatura, y tomaremos este valor como referencia de ahora en adelante.

4.1. Longitud de correlación

Otra propiedad fundamental de la transición de fase del modelo de Ising, que estudiaremos únicamente en el caso bidimensional, es la divergencia de ciertos observables que se da también en el límite termodinámico cuando alcanzamos T_c, como la susceptibilidad χ, el calor específico C_v o la longitud de correlación ξ. Condensaremos su estudio únicamente en la longitud de correlación, que caracteriza el tamaño de los clusters que forman los spines de mismo signo a diferentes temperaturas. Formalmente, definimos la correlación entre dos spines i, j cualesquiera

de la red según la expresión[3]:

$$C(i,j) = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \tag{7}$$

Este valor depende únicamente de la distancia entre spines, $C(i,j) = C(|\mathbf{r_i} - \mathbf{r_j}|) \equiv C(\mathbf{r_{ij}})$. Las correlaciones decaen exponencialmente con dicha distancia, lo que nos permite definir la longitud de correlación $\xi(T)$ según la expresión:

$$C(\mathbf{r_{ij}}) \propto e^{-\frac{|\mathbf{r_{ij}}|}{\xi(T)}} \tag{8}$$

Es decir, $\xi(T)$ marca la distancia en la que las correlaciones decaen un factor e. Buscaremos entrenar una red neuronal capaz de predecir $\xi(T)$ a partir de una configuración. Esta longitud de correlación diverge en T_c según $\xi(T) \propto |T - T_c|^{-\nu}$.

Dado que las correlaciones son únicamente función de la distancia entre spines, podemos realizar los promedios necesarios sobre todas las posibles parejas de spines en una configuración determinada. La importancia de este hecho radica en que partiendo de una única configuración, y aplicando el tratamiento matemático adecuado, obtenemos un único valor para la longitud de correlación, siempre el mismo[4]. Al entrenar la red neuronal nos aseguramos de que existe una relación directa, aunque compleja, entre la configuración y el output esperado.

Parte II
Resultados

5. Predicción de fase ordenada y desordenada

Nuestro primer objetivo es entrenar a la red para que, dada una configuración, sea capaz de distinguir en qué fase se encuentra. Esto se traduce en que la red debe tener una capa de entrada de $L \cdot L$ neuronas y una capa de salida con dos neuronas, una correspondiente a la fase ordenada y otra a la desordenada. Dada una configuración, la predicción de la red será aquella cuya neurona tome un mayor valor a la salida. Utilizamos una red con dos capas intermedias, la primera con 60 neuronas y la segunda con 30. Considerando un *bias* para cada neurona (excepto las del input) y un peso entre cada par de neuronas de capas adyacentes, tenemos un total de $L \cdot L \cdot 60 + 1952$ parámetros. En caso de una red típica con $L = 40$, trabajamos con 97952 parámetros.

[3] Utilizamos esta expresión y no únicamente $C(i,j) = \langle \sigma_i \sigma_j \rangle$ porque con esta última se producen divergencias en la longitud de correlación a baja temperatura, cuando todos los spines se orientan en la misma dirección. En este caso, $C(i,j) = 1 \ \forall i,j$, y por tanto $\xi(T) \to \infty$ por la expresión (8). En cambio, con nuestra ecuación, $C(i,j) = 0 \ \forall i,j$, porque a bajas temperaturas $\langle \sigma_i \sigma_j \rangle \simeq \langle \sigma_i \rangle \langle \sigma_j \rangle$, obteniendo finalmente $\xi(T) \to 0$.

[4] En realidad el cálculo de la longitud de correlación implica un ajuste exponencial según la relación (8), y por tanto su determinación conlleva cierto error inherente al método computacional utilizado para ello. En nuestro caso el ajuste se realiza mediante la función *optimize.curve_fit* de la librería *scipy* de *Python*.

5.1. Entrenamiento y predicción

Es importante entrenar a la red con una gran cantidad de configuraciones, y que éstas sean representativas del comportamiento que queremos enseñar. En nuestro caso esto se consigue tomando configuraciones en los límites $T \to 0$ y $T \to \infty$, donde sabemos con seguridad que son ordenadas y desordenadas, respectivamente. Si nos acercamos demasiado a la temperatura de transición en la búsqueda de configuraciones, se diluirán las características propias de estas fases y adulteraremos el entrenamiento, debido principalmente a que no podremos tener siquiera la seguridad de que la fase asignada sea la adecuada[5].

Utilizaremos para el entrenamiento configuraciones correspondientes a diferentes β en los rangos [0,1, 0,3] y [0,8, 1,0] a intervalos de $\Delta_\beta = 0{,}02$. Se encuentran por tanto muy alejados del valor crítico $\beta_c \simeq 0{,}44$. Para cada temperatura generaremos 1000 configuraciones, dando lugar finalmente a un total de 20000. Como hemos comentado anteriormente, debemos asegurarnos de que en la fase ordenada aparecen configuraciones con los spines apuntando en ambas direcciones. Para ello, recorreremos el rango de temperaturas dos veces con fases ordenadas apuntando en dirección contraria, dando lugar a 40000 configuraciones.

Con estos datos podemos pasar a entrenar la red aplicando los procedimientos descritos en la primera sección. Como hiperparámetros tomamos $\eta = 0{,}5$ para cada paso de actualización de los parámetros por descenso de gradiente, y $\lambda = 10$ para la regularización L2. Tomaremos un tamaño de *mini-batch* de 10 configuraciones. Estas elecciones, si bien no son completamente arbitrarias, tienen poca influencia en el rendimiento de la red; otras elecciones de los parámetros no producen cambios sensibles ni en el resultado ni en la velocidad. Esto se debe principalmente a la simpleza del modelo, que da lugar a convergencias muy rápidas del proceso logrando una precisión del 100 % de aciertos para los datos de entrenamiento en unas pocas épocas, en ningún caso más de diez. Para evitar el overfitting, hemos tomado como criterio que el entrenamiento termine tras la segunda época consecutiva en la que la red clasifique correctamente todas las configuraciones de entrenamiento. De esta forma aseguramos que es capaz de realizar las predicciones, sin dar tiempo a sobreajustar la red.

Una vez entrenada la red, necesitamos una gran cantidad de configuraciones en torno a β_c para llevar a cabo la predicción. En general utilizaremos valores de β entre [0,4, 0,45], con $\Delta_\beta = 0{,}002$. Para cada temperatura generamos 1000 configuraciones, dando lugar a un total 25000.l Al igual que en el caso anterior, duplicamos esta cantidad para obtener 50000. Con estas configuraciones realizamos la predicción de la siguiente manera: introducimos en la red todas aquellas configuraciones correspondientes a la misma temperatura, y realizamos un promedio con todas las predicciones de cada neurona por separado[6]. Repitiendo este procedimiento para todas las temperaturas podemos representar una gráfica con dos series de datos correspondientes a ambas predicciones. Asignaremos la transición de fase al punto de corte entre ambas series.

[5] En un sistema finito no existe una frontera nítida para la transición, sino que en cierto rango de temperaturas aparecen configuraciones muy difíciles de clasificar, que dan lugar a magnetizaciones medias pequeñas pero diferentes de 0 para temperaturas mayores que la crítica. Por tanto, asignar como único criterio para la clasificación que la configuración haya aparecido a una β mayor o menor que la crítica no resulta válido.

[6] Es decir, promediamos por un lado las predicciones de la neurona correspondiente a la fase desordenada, y por otro las de la desordenada.

5.2. Redes *fully connected*

Se muestran en la figura (3) los resultados para una red de Ising ferromagnética con $L = 40$, que sitúan la transición de fase en $\beta_c = 0{,}426$. Vemos que en un rango relativamente amplio de β conviven configuraciones ordenadas y desordenadas, según la red neuronal. Estos resultados tienen bastante variabilidad en la tercera cifra decimal para diferentes entrenamientos de la red. Se observan dos líneas casi superpuestas porque cada una de ellas corresponde a una orientación de spines diferente en la fase ordenada. Vemos que prácticamente no hay efectos de histéresis y el sistema es capaz de aprender las fases independientemente de la orientación de los spines.

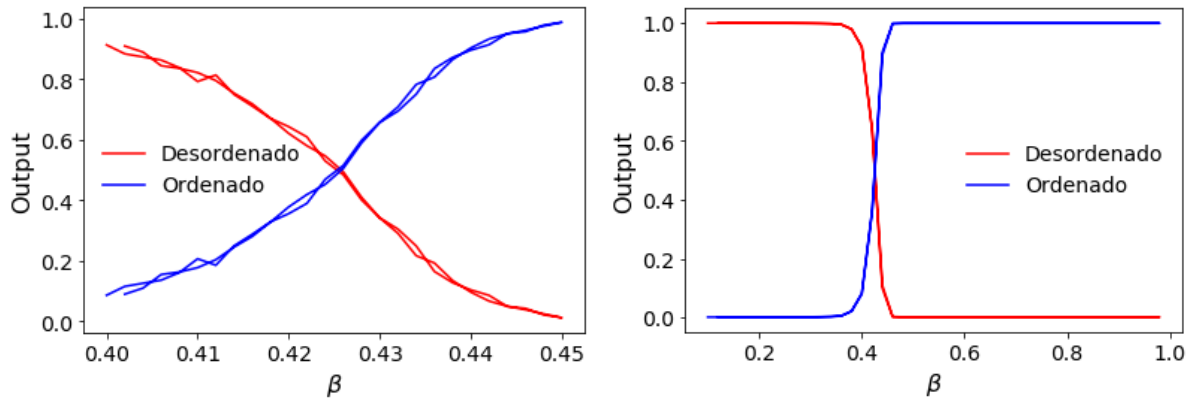


Figura 3: Resultados para las predicciones del modelo de Ising ferromagnético con $L = 40$ haciendo uso de una red *fully connected*. Se muestra el output promedio de cada neurona por separado frente a β. (a) Utilizamos las configuraciones generadas con $\beta \in [0,4,\ 0,45]$ para predecir con precisión la transición de fase, que se sitúa en $\beta_c = 0,426$. (b) Mostramos una imagen general de las predicciones de la red en el rango $\beta \in$ [0,1, 1,0] con $\Delta_\beta = 0{,}05$, para poner en contexto las gráfica en (a) y mostrar la precisión de las predicciones.

Si bien el valor teórico se sitúa en 0,4406, hay que tener en cuenta que para una red de tamaño finito no existe una frontera clara entre fases y la magnetización promedio deja de ser nula para valores de β más pequeños. Se trata por tanto de un comportamiento esperable, que se debería corregir para redes de mayor tamaño. Realizamos ahora el mismo experimento con redes de Ising de diversos tamaños; los resultados se muestran en la figura (4). Vemos que existe poca variabilidad en la determinación de diferentes valores de β_c para una misma L dando lugar a errores muy pequeños en el ajuste, y que aparece una clara tendencia ascendente conforme nos acercamos al límite termodinámico. Realizando un fit lineal obtenemos $\beta_c = 0{,}4362 \pm 0{,}0012$ cuando $L \to \infty$, que se desvía del valor teórico en poco más de un 1 %. Se observa pues que se reproduce el comportamiento esperado con L.

En definitiva, vemos que la red es capaz de predecir correctamente la transición de fase. Podemos indagar ahora en su funcionamiento interno para intentar explicar este comportamiento. Existen dos características principales en las que se puede fijar la red para realizar sus predicciones. Por un lado, puede estar aprendiendo la magnetización de la red de Ising y prediciendo a partir de ésta. Por otro, puede codificar también las relaciones entre spines que caracterizan el orden del modelo y también determinan la transición de fase. Centrándonos en la magnetización, vemos en la figura (5) que no es el único criterio para determinar las predicciones de la

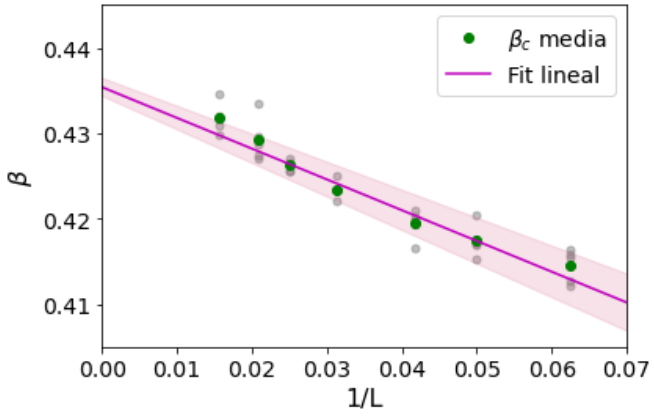


Figura 4: Se representan las predicciones de la transición de fase frente a $1/L$, lo que nos permite extrapolar las tendencias al límite termodinámico, cuando $1/L = 0$. Se han utilizado redes con $L = 64$, 48, 40, 32, 24, 20 y 16. Los valores en verde muestran la β_c promedio, obtenida a partir de cinco entrenamientos diferentes de la red para cada L, que se muestran en gris. La desviación estándar no se muestra en la gráfica porque en la mayoría de ocasiones es más pequeña que el tamaño del punto. El ajuste lineal arroja un valor en el límite termodinámico de $\beta_c = 0{,}4362 \pm 0{,}0012$.

red neuronal, aunque ambas estén claramente relacionadas. Para ello basta fijarse en la anchura de la parte de transición, indicando que la red asigna las mismas predicciones para redes con magnetizaciones que difieren hasta en $0, 2$.

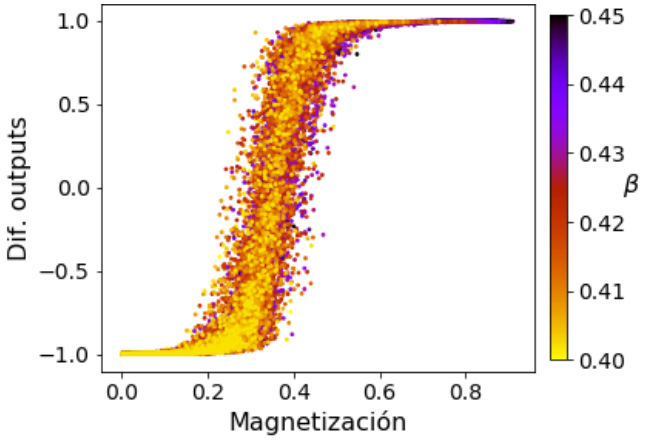
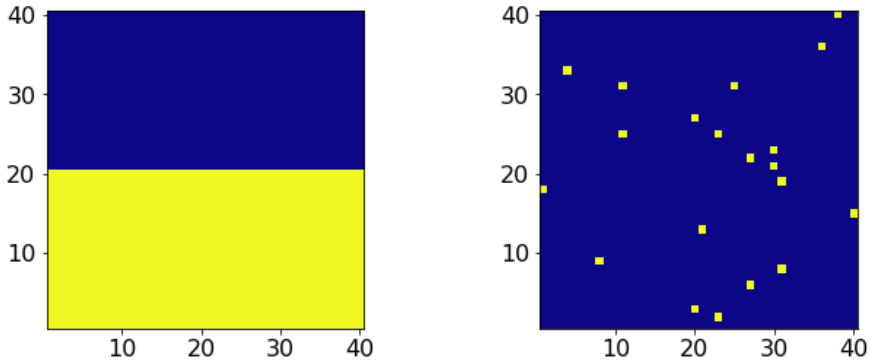


Figura 5: En el eje vertical se representa la diferencia de outputs entre la neurona ordenada y desordenada frente a la magnetización de la configuración. Por tanto, todo valor superior a 0 indica que la configuración se considera ordenada, y menor que 0, desordenada. Se observa inmediatamente que el proceso no es un simple aprendizaje de la magnetización del sistema.

Este comportamiento tiene un problema fundamental, porque pueden existir configuraciones con magnetización nula que por su energía puedan corresponder a fase ordenada. Vemos un ejemplo en la figura (6a). Recordemos que la energía es mayor cuantas más fronteras entre spines de distinto signo hay, y a mayor energía, menor probabilidad de aparecer a temperaturas bajas. Por tanto, es razonable asumir que una configuración con energía muy baja se debería considerar correspondiente a fase ordenada, donde la probabilidad de que aparezca es infinitamente

mayor que en la fase desordenada al competir con una cantidad mucho menor de configuraciones. Sin embargo, esto no ocurre para las configuraciones mostradas: mientras que la segunda configuración se asigna correctamente a la fase ordenada, la primera se asigna a la desordenada. Esto apunta a una fuerte dependencia con la magnetización en redes del tipo *fully connected*. Si bien no es un comportamiento que podamos calificar como incorrecto, siembra dudas sobre otras predicciones con casos similares, pese a que no representa un gran problema dada la baja probabilidad de este tipo de configuraciones. Estos resultados también apuntan a que, en caso de existir, el conocimiento de la topología de la red se da de forma global; si la red neuronal valorase el orden general de la configuración según el orden local en pequeñas zonas de la misma (grandes conjuntos de spines correctamente ordenados que dan lugar a magnetizaciones no nulas en amplias zonas de la red), la configuración (6a) se debería considerar ordenada. En cualquier caso este comportamiento es esperable debido a que las configuraciones para el entrenamiento se generan en temperaturas en las que el orden local no es tan importante[7], y por tanto en principio la red no es capaz de aplicar este criterio.



(a) (b)

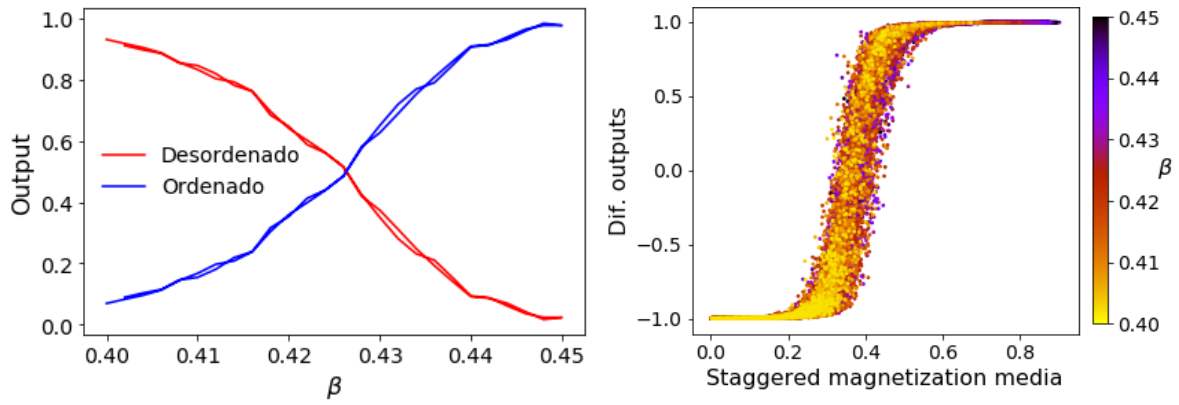
Figura 6: Configuraciones del modelo de Ising con $L = 40$. Cada color corresponde a una orientación de spin. Ambas configuraciones tienen la misma energía, correspondiente a 80 fronteras entre spines de diferente signo (condiciones de contorno periódicas). Sin embargo, la configuración (a) presenta una magnetización global nula, mientras que la configuración (b) tiene una magnetización muy cercana a 1.

5.2.1. Relaciones entre spines. Ising antiferromagnético

Para estudiar cómo aprende la red las relaciones entre spines resulta muy útil analizar el comportamiento con un sistema antiferromagnético. Sabemos que también existe una transición de fase que no depende del signo de J y por tanto es la misma que en el sistema ferromagnético. En este caso la magnetización es prácticamente nula para todo el espectro de temperaturas, y como hemos visto el parámetro de orden más adecuado es la *staggered magnetization*, la magnetización de las subredes del sistema. Centraremos el estudio en las propiedades de una red de Ising antiferromagnética con $L = 40$, utilizando un método completamente análogo al caso ferromagnético. Los resultados obtenidos se muestran en la figura (7). Se comprueba que el

[7]Estos grandes conjuntos de spines ordenados son propios de las regiones de temperaturas cercanas a la transición de fase, en las que la longitud de correlación diverge y las fluctuaciones aumentan mucho su tamaño.

sistema es capaz de aprender a partir de los datos de entrenamiento y obtener una predicción para la transición de fase, que se sitúa en $\beta_c = 0{,}426$, el mismo valor que en el caso ferromagnético. Vemos también que la relación entre las predicciones y la media de *staggered magnetizations* de ambas subredes es muy similar al caso anterior, situándose la transición en unos valores parecidos de la magnetización y con similar anchura.



(a) (b)

Figura 7: Resultados para las predicciones del modelo de Ising antiferromagnético con $L = 40$ haciendo uso de una red *fully connected*. (a) Predicción de la transición de fase en función de β, que se da en $\beta = 0,426$. (b) Diferencia entre outputs ordenado y desordenado en función de la media de las *staggered magnetizations* de ambas subredes para cada configuración.

Es importante notar que el cálculo de la *staggered magnetization* no es sencillo para la red: debe promediar por separado aquellos spines cuya suma de ambos términos $i + j$ de sus coordenadas (i, j) dé lugar a números pares e impares. A esto nos referimos cuando hablamos de la topología de la red, a ser capaz de determinar la combinación de spines que hace aflorar el orden correcto del modelo. Aunque la gráfica (7b) parece indicar que este conocimiento sí se da al lograr determinar correctamente la *staggered magnetization*, es posible (aunque improbable) que sea consecuencia únicamente de la aplicación de un criterio rudimentario de elevada alternancia entre spines, lo que podría dar lugar a esta correlación entre predicciones y magnetizaciones sin que la red llegue a establecerla como criterio.

Para profundizar en el funcionamiento de la red, utilizamos las configuraciones mostradas en (8). La ventaja de utilizar el modelo antiferromagnético reside en que la fase ordenada tiene una estructura característica de tablero de ajedrez, como se muestra en la figura (8a). Cuando consideramos este tipo de estructuras como un único vector, que es lo que se introduce en la red neuronal, nos encontramos con una sucesión de spines cuyas direcciones se van alternando casi a la perfección[8]. En caso de que la red no fuera capaz de aprender las relaciones verticales entre los spines, toda configuración con una elevada alternancia de spines se debería considerar ordenada, pues se parecen mucho a la configuración completamente ordenada. La configuración (8b) muestra una alternancia perfecta de spines al construir el vector, pero vemos claramente que la configuración en dos dimensiones no corresponde a una fase ordenada, sino que en vertical

[8]En realidad son completamente alternos salvo en las uniones entre dos filas, en las que se juntan dos spines que apuntan en la misma dirección.

todos los spines apuntan en una misma dirección y dan lugar a un alto valor de la energía. Además, la *staggered magnetization* es nula en ambas subredes, por lo que también según este criterio la predicción debería corresponder a fase desordenada. En efecto, la red acierta en su predicción y clasifica correctamente la configuración como desordenada, indicando que la simple alternancia entre spines no determina completamente las predicciones de la red, sino que se asignan más criterios.

Por último, la configuración (8c) corresponde a una fase ordenada en la que la sección inferior se ha desplazado una spin hacia la derecha, dando lugar a una configuración cualitativamente análoga a la mostrada en (6b) para el caso ferromagnético. Al igual que antes, presenta por un lado una *staggered magnetization* nula, lo que podría indicar una fase desordenada, pero por otro tiene una energía muy pequeña, propia de las configuraciones ordenadas; ambas propiedades compiten. La predicción para esta configuración corresponde a una fase desordenada, lo que indica que la red es capaz de calcular la *staggered magnetization*, con las dificultades que esto conlleva, y además la aplica como criterio principal para sus predicciones. En este caso el uso de la *staggered magnetization* conlleva necesariamente cierto conocimiento de la topología de la red de spines para determinar cuáles de estos deben sumarse entre sí. No obstante, al igual que en el caso ferromagnético, este conocimiento de las magnetizaciones es global y no parece valorar las relaciones entre spines adyacentes, únicamente es capaz de separarlos en dos grandes grupos para realizar la suma, es un conocimiento global.

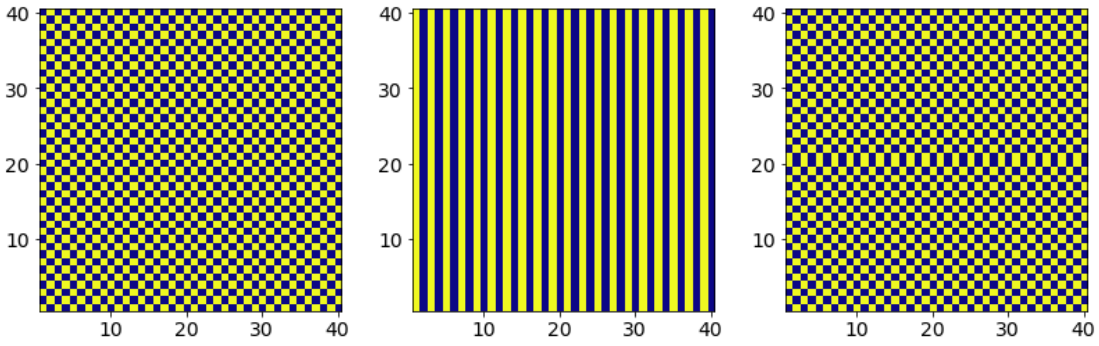
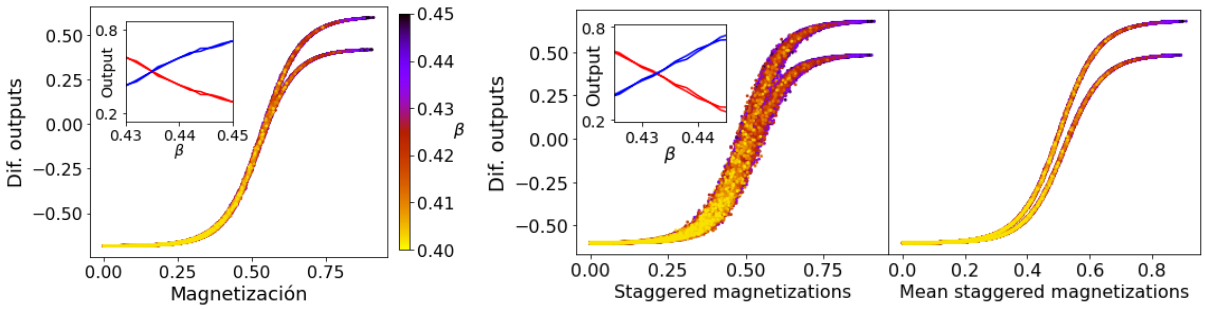


Figura 8: Configuraciones para el modelo de Ising antiferromagnético bidimensional. (a) Configuración completamente ordenada. (b) Configuración ordenada en la que las filas pares se han desplazado una unidad hacia un lado, de forma que cada spin ya no está rodeado únicamente de spines contrarios, sino que dos de sus vecinos tienen su misma orientación. Nótese que esto da lugar a una *staggered magnetization* nula para ambas subredes. (c) Configuración ordenada, en la que la mitad inferior se ha desplazado un spin hacia la derecha. Al igual que en (b) la *staggered magnetization* es nula en ambas subredes, pero la configuración tiene una energía mucho menor.

5.2.2. Disminución de los parámetros. Redes estranguladas

Podemos realizar un experimento interesante consistente en estudiar qué ocurre cuando trabajamos con redes extremadamente simples, con muy pocos parámetros. De esta forma, hacemos aflorar las características más simples del modelo que nos permiten realizar las predicciones. En

este apartado construimos una red neuronal con una única capa intermedia de 3 neuronas (en total, 4811 parámetros) y llevamos a cabo el mismo proceso de entrenamiento que en el caso anterior para una red de Ising con $L = 40$. Para asegurar que trabajamos con el menor número de parámetros posibles, utilizamos un hiperparámetro de regularización extremadamente grande, $\lambda = 1000$. Los resultados para el caso ferromagnético se muestran en la figura (9a). Vemos que somos capaces de predecir una transición, pero basada únicamente en la magnetización de la configuración, como indica la anchura reducida de la línea formada por las predicciones. Se observa además que para diferentes orientaciones del spin en la fase ordenada el output es sensiblemente diferente, lo que indica que la red ni siquiera es capaz de generalizar su comportamiento a ambas orientaciones, aunque las predicciones finales sean correctas. Por último, en línea con los apartados anteriores, la red neuronal predice incorrectamente que configuraciones con magnetizaciones de en torno a $0,5$ corresponden a la fase desordenada.



(a) (b)

Figura 9: Predicciones de la red neuronal simple para el modelo de Ising con $L = 40$. (a) Caso ferromagnético. Diferencia entre *outputs* de las neuronas ordenada y desordenada frente a la magnetización de la configuración. Inset: predicción para la transición de fase, que se sitúa en $\beta_c = 0{,}435$. (b) Caso antiferromagnético. A la izquierda se muestran las predicciones frente a las *staggered magnetizations* de cada configuración. Aparecen por tanto dos puntos por configuración. Inset: predicción para la transición de fase, que se sitúa en $\beta_c = 0{,}432$. A la derecha se muestran las predicciones frente a la media de *staggered magnetizations* de cada configuración.

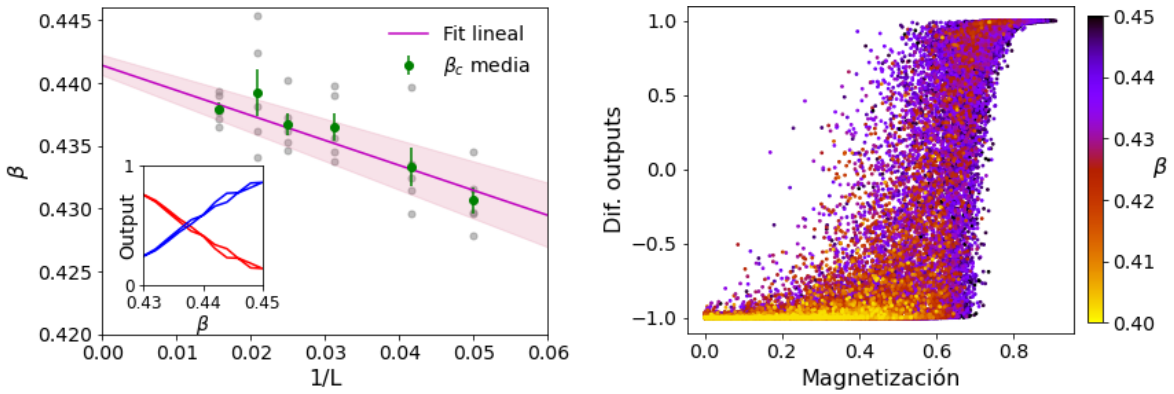
Podemos realizar un procedimiento igual al anterior con el caso antiferromagnético. Utilizando los mismos parámetros, obtenemos los resultados que se muestran en la figura (9b). Si bien son similares a los del caso ferromagnético, destaca que en esta ocasión las predicciones dependen claramente del valor medio de las *staggered magnetizations* de cada configuración, lo que requiere cierto conocimiento de la topología de la red para calcular correctamente las magnetizaciones de cada subred, y después un cálculo adicional para determinar la media. Este cálculo no es sencillo, y ejemplifica la capacidad de las redes neuronales *fully connected*, que incluso con muy pocos parámetros son capaces de obtener información topológica acerca de la estructura general de los datos.

5.3. Redes convolucionales

Realizaremos ahora el mismo experimento con redes convolucionales. Como hemos comentado anteriormente, esto nos permite estudiar las configuraciones desde un punto de vista local,

es decir, fomentando la búsqueda de relaciones entre spines cercanos que permitan caracterizar el orden de la red de Ising a un nivel global, aprovechando que se conoce la topología de la configuración. Esto implica que la magnetización total de la configuración ya no debería ser el principal indicador del orden de la red, porque ésta no es capaz de calcularla de un modo global. En pocas palabras, la red debe aprender si el sistema está o no ordenado a partir de la presencia o ausencia de orden en zonas locales.

Para construir las redes convolucionales utilizamos *Tensorflow*, que implementa una gran cantidad de herramientas propias del aprendizaje automático en una librería de *Python*. Los fundamentos básicos de las funciones incluidas son los mismos que los explicados hasta ahora, y el funcionamiento de la red es completamente análogo al caso anterior. En este apartado utilizaremos una red con dos capas convolucionales intermedias, cada una de ellas con cuatro filtros y un *receptive field* de $(3, 3)$, compuestas por neuronas *Leaky ReLU*. A continuación introducimos una *average pooling layer* que promedia los *outputs* de la capa anterior con un *receptive field* de $(2, 2)$, lo que reduce el tamaño de las matrices a la mitad. Añadimos después una capa *fully connected* de 30 neuronas *ReLU*, y finalmente obtenemos el *output* de una capa con dos neuronas *softmax*. En total, para una red de Ising con $L = 40$ tenemos 48280 parámetros, aproximadamente la mitad que en el caso anterior. Respecto al resto de hiperparámetros, no incluimos regularización[9], y utilizamos un tamaño de *mini-batch* de 32 configuraciones.



(a) (b)

Figura 10: Resultados para las predicciones del modelo de Ising ferromagnético haciendo uso de una red convolucional. (a) Predicciones de la transición de fase frente a $1/L$. Se han utilizado redes con $L = 64$, 48, 40, 32, 24 y 20. El ajuste lineal arroja un valor en el límite termodinámico de $\beta_c = 0{,}4414 \pm 0{,}0008$. Se muestra en el inset una de las predicciones con $L = 40$, dando lugar a $\beta_c = 0{,}437$. (b) Diferencia de *outputs* ordenado y desordenado frente a la magnetización de la configuración para una red de Ising con $L = 40$.

Mostramos en la figura (10a) los resultados obtenidos para una red de Ising con diferentes valores de L. Cabe resaltar que no hemos incluido resultados con $L = 16$, puesto que el diseño de la red neuronal da lugar a un total de 7960 parámetros, demasiado pocos como para obtener resultados válidos[10]. Observamos que la predicción de fase se realiza de forma correcta en todos

[9]Las redes convolucionales muestran una notable resistencia al *overfitting* por su propia naturaleza, que promedia su comportamiento sobre toda la red. Además, partimos de un reducido número de parámetros.

[10]En principio nada impide utilizar redes con distinto número de neuronas y capas, puesto que el número de

los casos, obteniendo un valor en el límite termodinámico de $\beta_c = 0,4414 \pm 0{,}0008$, que incluye el valor teórico de $\beta_c^{teor} = 0{,}4406$ dentro del intervalo de confianza. Se trata por tanto de un resultado muy satisfactorio.

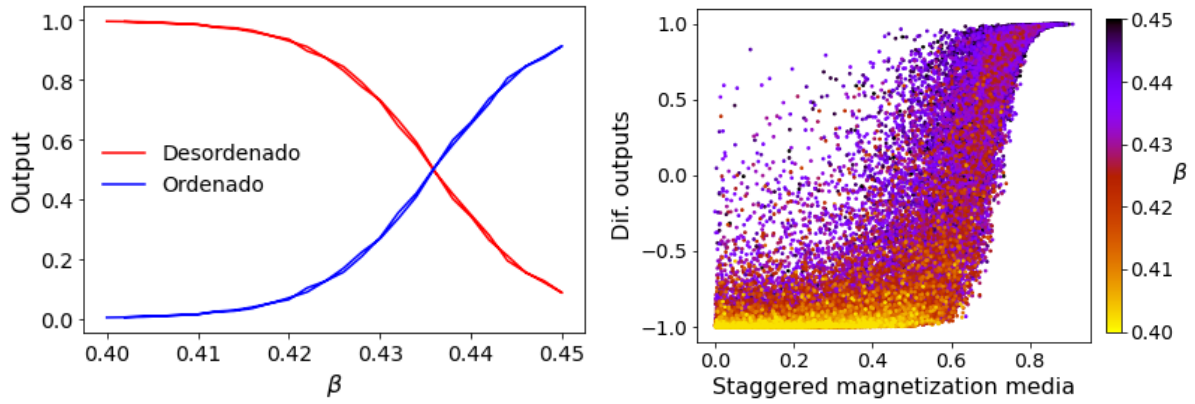
La diferencia principal frente a los resultados anteriores aparece en la imagen (10b), donde se ensancha la zona de transición mostrando que la magnetización pierde importancia para las predicciones. Destaca la gran cantidad de configuraciones con magnetizaciones elevadas, por encima de 0,6, que se predicen como desordenadas. Es un umbral bastante mayor que el obtenido para redes *fully connected* (ver figura (7)), lo que explica el desplazamiento de la beta crítica frente a resultados anteriores. Es difícil justificar que configuraciones con magnetizaciones tan elevadas se clasifiquen como desordenadas (pese a que esto puede ocurrir durante las simulaciones, recordemos que en fase desordenada todas las configuraciones son equiprobables), y podemos considerarlo un mal comportamiento, que se corrige no obstante conforme aumenta L al hacerse más abrupta la transición. Podíamos esperar este resultado, puesto que por su estructura la red convolucional es incapaz de utilizar observables que se calculan de forma global. Vemos además por la anchura de la franja de transición que la magnetización deja de ser un factor determinante en la predicción final. Aparecen también algunas configuraciones con valores muy bajos de la magnetización que se consideran ordenadas, una propiedad que permite a la red clasificar correctamente las configuraciones mostradas en la figura (6), situándolas ambas en la fase ordenada pese a valor nulo de la magnetización de la primera de ellas. En definitiva, el comportamiento de las redes convolucionales resulta en cierto modo complementario al anterior, centrándose en las relaciones locales y no en los parámetros de orden.

Lo mismo ocurre en el caso antiferromagnético, que mostramos en la figura (11). Observamos que la transición de fase se predice correctamente en $\beta_c = 0{,}436$, muy cercano al valor real y a la predicción realizada en el caso ferromagnético. Llama la atención también que las predicciones parecen tener mucho menos en cuenta la *staggered magnetization* que la magnetización en el caso ferromagnético, ensanchando mucho la zona de transición en la gráfica (11b), aunque el comportamiento es cualitativamente similar clasificando configuraciones con *staggered magnetizations* por encima de $0,7$ como fase desordenada. Esto puede deberse a que la *staggered magnetization* es más complicada de aprender para la red, por lo que la podría tener menos en cuenta. Por último, esta red neuronal clasifica correctamente todas las configuraciones mostradas en la figura (8), de manera similar al caso ferromagnético con sus configuraciones.

6. Longitud de correlación

Trataremos de entrenar ahora a la red en la predicción de la longitud de correlación ξ del modelo de Ising. El objetivo es determinar el valor de β para el que se da la máxima longitud de correlación, que marca la transición de fase β_c. A diferencia del caso anterior, en el que llevábamos a cabo un proceso de clasificación, en esta ocasión necesitamos obtener un valor numérico determinado en el *output*. Esto se traduce en que la última capa está formada únicamente por

parámetros no se mantiene constante para distintos valores de L, y se podrían justificar cambios en la red neuronal para aumentar su número en caso de ser demasiado pocos. No obstante, hemos preferido limitarnos a los casos en los que podemos mantener la estructura inicial.



(a) (b)

Figura 11: Resultados para las predicciones del modelo de Ising antiferromagnético con $L = 40$ haciendo uso de una red convolucional. (a) Predicción de la transición de fase, que se sitúa en $\beta_c = 0{,}436$. (b) Diferencia de *outputs* ordenado y desordenado frente a la magnetización de la configuración.

una neurona. Además, la longitud de correlación puede tomar cualquier valor positivo, y no se limita al rango entre 0 y 1, por lo que en caso de utilizar neuronas en la última capa cuya salida se encuentre limitada como la sigmoide, necesitaremos llevar a cabo un proceso de normalización de los *outputs* antes del entrenamiento y predicción. En nuestro caso el criterio será asignar la unidad al mayor valor de la longitud de correlación entre todas las configuraciones de entrenamiento y predicción y reescalar el resto de longitudes frente a esta. La desventaja principal es que necesitaremos tener siempre presente el valor máximo con el que hemos normalizado para reescalar el resultado de las predicciones, y que seremos incapaces de determinar correctamente las longitudes de correlación por encima de dicho valor. La solución evidente pasa por utilizar neuronas a la salida con el *output* ilimitado, para lograr un comportamiento universal.

6.1. Entrenamiento y predicción

Como hemos comentado anteriormente, a cada configuración le corresponde una única longitud de correlación que se obtiene calculando las correlaciones para todas las parejas de spines (formadas por un spin y cualquiera de sus vecinos en direcciones vertical y horizontal) según la relación (7), y ajustando la exponencial según la expresión (8). En el proceso de entrenamiento necesitamos utilizar configuraciones cuyas longitudes de correlación abarquen rangos amplios, para que la red sea capaz de aprender correctamente en todo rango de temperaturas y sus predicciones resulten válidas incluso cerca de la temperatura crítica, en la que la longitud de correlación diverge. Para lograr esto, utilizamos configuraciones obtenidas en el rango $\beta_c \in [0{,}1,\ 0{,}8]$, con $\Delta\beta = 0{,}01$, atravesando claramente β_c y abarcando al mismo tiempo las fases ordenadas y desordenadas. Generamos 100 configuraciones para cada valor de β, pero dado que en muchas configuraciones alejadas de la zona de transición la longitud de correlación es prácticamente nula, generaremos 1000 configuraciones para cada β en el rango $[0{,}35,\ 0{,}5]$. De esta manera, garantizamos un número suficiente de configuraciones con longitudes de correlación más abultadas. Al igual que en el caso anterior incluiremos configuraciones con fases ordenadas

apuntando en direcciones opuestas, dando lugar a un total de 41000 configuraciones.

Centraremos el proceso de predicción en el intervalo $\beta_c \in [0{,}3,\ 0{,}6]$, con $\Delta\beta = 0{,}002$. Utilizaremos 500 configuraciones para cada valor de β, y recorreremos el intervalo dos veces según la orientación de la fase ordenada, dando lugar a un total de 150000 configuraciones. Se muestra en la figura (12a) la longitud de correlación en función de β obtenida como promedio de longitudes calculadas para todas las configuraciones correspondientes a una misma β.

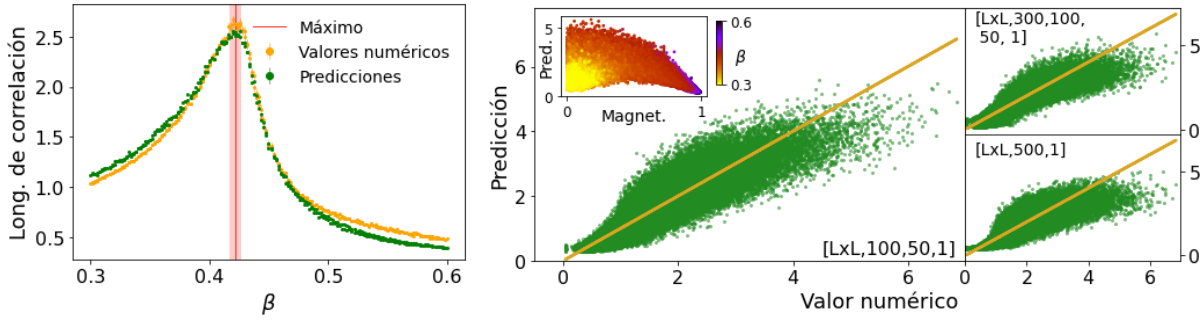
6.2. Redes *fully connected*

Utilizamos de nuevo una red neuronal con dos capas intermedias, pero en esta ocasión de 100 y 50 neuronas sigmoides respectivamente. El resto de hiperparámetros, salvo el de regularización que toma un valor $\lambda = 0{,}01$, se mantienen igual que en el caso de la predicción de fase. De nuevo, esta elección de parámetros no es única, existen otras combinaciones que dan lugar a buenos comportamientos de la red. A la salida situamos una única neurona sigmoide, puesto que esperamos un resultado entre 0 y 1 para cada configuración, siendo 1 el valor asignado para la mayor longitud de correlación entre todas las configuraciones de entrenamiento y predicción.

Un problema importante que aparece en esta ocasión es que no existe un criterio claro para terminar el proceso de entrenamiento, porque a diferencia de las redes utilizadas en clasificación en las que podemos determinar de forma clara los aciertos en las predicciones, en este caso no tenemos una métrica que nos indique si la red está funcionando correctamente. Únicamente podemos fijarnos en el valor que adquiere la función coste tras cada época de entrenamiento y tratar de minimizarlo. Además, este valor no disminuye de forma uniforme, sino que puede mantenerse constante o incluso aumentar en algunas épocas antes de seguir disminuyendo. Por tanto, resulta complicado implementar un criterio de *early stopping* que funcione en todos los casos. Para superar esta dificultad, seleccionamos aleatoriamente un 5 % de las configuraciones de entrenamiento y las apartamos, asegurando que la red no las reciba como *input* y no las pueda utilizar en el aprendizaje. Monitorizamos el valor de la función coste sobre estas configuraciones, de forma que podamos tener una imagen razonablemente buena de la capacidad de la red para predecir correctamente la longitud de correlación en configuraciones que no ha visto anteriormente. A partir de esto, establecemos como criterio para terminar el entrenamiento que el valor que adquiera la función coste en una época determinada sea mayor que la media de las últimas veinte épocas. Además, establecemos un límite máximo de cincuenta épocas de entrenamiento.

Mostramos en la figura (12a) las predicciones para una red de Ising ferromagnética con $L = 40$ junto con los resultados numéricos. En la figura (12b) se muestran las predicciones obtenidas por la red en función del valor numérico para cada configuración. La diagonal, marcada en naranja, corresponde al funcionamiento ideal de la red neuronal, en el que las predicciones son iguales que el resultado obtenido numéricamente. Lo primero que destaca en esta gráfica es que las predicciones parecen seguir una tendencia general acertada, pero se desvían mucho cuantitativamente dando lugar a una nube de predicciones para valores intermedios. Observamos también que un aumento significativo del número de parámetros de la red no parece mejorar los resultados, y podemos asumir que el comportamiento mostrado es el esperable y representativo de redes neuronales *fully connected*. En cualquier caso, la línea diagonal parece situarse en un punto intermedio entre las predicciones para buena parte del rango de longitudes de correlación,

por lo que es posible que los errores de unas predicciones y otras para una misma β se compensen dando lugar a un buen ajuste. En efecto, esto lo observamos en la gráfica (12a), en la que pese a los malos resultados para cada configuración por separado, obtenemos promedios que se ajustan muy bien al valor real. Es aún más importante para nosotros el hecho de que las predicciones acierten en la β en la que se produce el máximo de la longitud de correlación, pues esto determina la transición de fase.



(a) (b)

Figura 12: Predicciones para una red de Ising ferromagnética con $L = 40$ haciendo uso de una red neuronal *fully connected*. (a) Predicciones y resultados numéricos para el valor medio de la longitud de correlación en función de β. El máximo de las predicciones se sitúa en $\beta_c = 0{,}422 \pm 0{,}04$. (b) En verde, predicción para la longitud de correlación frente al valor obtenido de forma numérica. La línea dorada marca la diagonal, el valor que deberían tomar las predicciones. Todas las predicciones han sido reescaladas para representar el valor real de la longitud de correlación. Izquierda: Resultados obtenidos para la red original, con estructura [LxL, 100, 50, 1]. Derecha: resultados obtenidos para diversas redes con un mayor número de parámetros. Inset: predicciones de la red en función de la magnetización de la configuración.

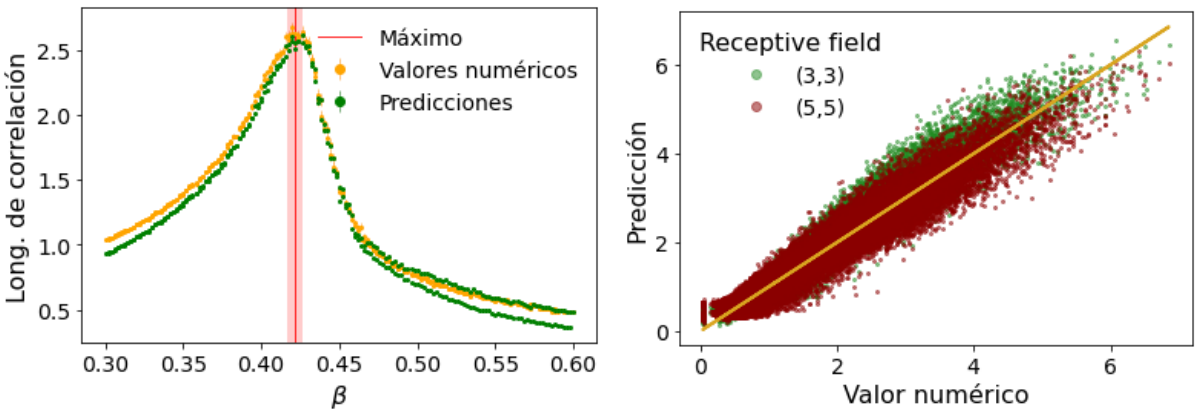
También se aprecia que la red parece tener un umbral inferior para las predicciones, y no es capaz de alcanzar el valor nulo debido a que necesita *inputs* muy negativos para ello, difíciles de conseguir durante el entrenamiento. Se trata por tanto de un problema de construcción de la red, que se puede solucionar sustituyendo la función de la neurona a la salida. Por otro lado, la red parece infravalorar también las longitudes de correlación más elevadas. No obstante, esto no corresponde al mismo efecto de saturación que para los valores más bajos, puesto que el valor máximo de las correlaciones con el que se normaliza el *output* se sitúa en torno a $7,5$ y corresponde a una configuración de entrenamiento, lejos del mayor valor que encontramos en las configuraciones para la predicción. Por tanto, la desviación de los datos probablemente se deba a una falta de configuraciones con longitudes de correlación elevadas en el conjunto de entrenamiento, evitando un correcto aprendizaje. La incapacidad de este tipo de redes para predecir correctamente el valor de la longitud de correlación es esperable teniendo en cuenta que la topología de la red se conoce de forma general. Vemos además en el inset de la figura (12) que no existe una relación clara entre la predicción de la longitud de correlación y la magnetización, que deja de usarse como criterio.

En cualquier caso, el éxito principal de la red neuronal consiste en ser capaz de determinar correctamente la β a la que se da la máxima longitud de correlación, pues en ese punto se sitúa la transición de fase. En este sentido, vemos que en efecto los máximos obtenidos de forma

numérica y mediante la red neuronal coinciden en $\beta_c = 0{,}422 \pm 0{,}04$. El límite para la precisión corresponde a la anchura del máximo calculado de forma numérica, factor limitante en esta predicción. Se trata no obstante de un resultado muy satisfactorio.

6.3. Redes convolucionales

Realizamos ahora el mismo procedimiento, haciendo uso esta vez de redes neuronales convolucionales. En esta ocasión utilizamos una red formada por dos capas convolucionales con 8 filtros cada una y un *receptive field* de (5,5), formadas por neuronas *leaky ReLU*, una capa *fully connected* con 30 neuronas *ReLU*, y una capa de salida con una única neurona sigmoide a la salida. El aumento de los filtros en las capas convolucionales y el tamaño del *receptive field* responde al aumento de dificultad de las predicciones. Los resultados para el promedio de las predicciones se muestran en la figura (13a). Como vemos, los promedios de las predicciones se ajustan correctamente a los valores numéricos, y la posición del máximo es la adecuada, en $\beta_c = 0{,}422 \pm 0{,}04$. Se trata del mismo resultado que en el caso anterior, al estar limitada la precisión por los propios datos de entrenamiento y no por el desempeño de la red neuronal. Destacan los problemas de la red para asignar correctamente la longitud de correlación según la orientación del spin en la fase ordenada, un problema mucho menos frecuente en la red anterior.



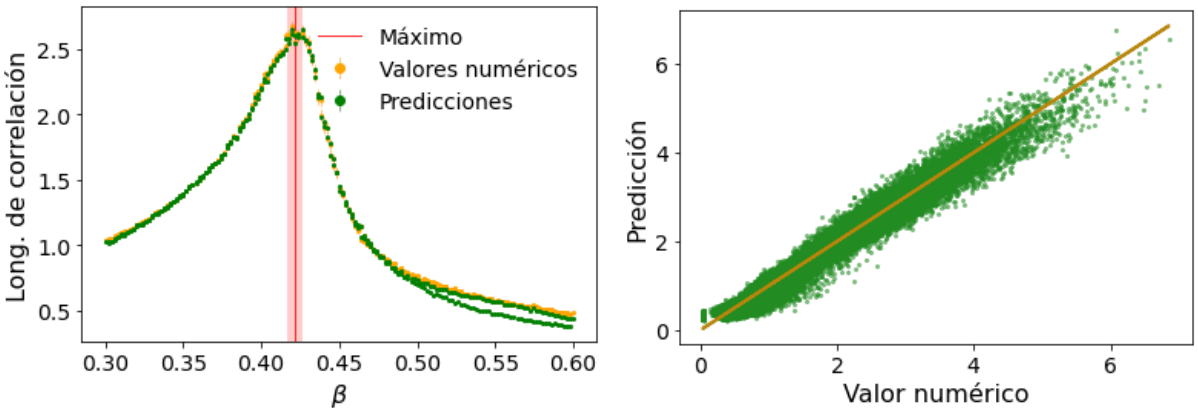
(a) (b)

Figura 13: Predicciones para una red de Ising ferromagnética con $L = 40$ haciendo uso de una red convolucional con una neurona *sigmoid* a la salida. (a) Predicciones y resultados numéricos para el valor medio de la longitud de correlación en función de β. El máximo se sitúa de nuevo en $\beta_c = 0{,}422 \pm 0{,}04$. (b) Predicciones para la longitud de correlación que obtenemos de la red frente al valor obtenido de forma numérica. Todas las predicciones de la red han sido reescaladas para representar el valor real de la longitud de correlación. La línea dorada marca la diagonal, el valor que deberían tomar las predicciones. En verde, resultados obtenidos con un *receptive field* de $(3, 3)$. En rojo, resultados obtenidos con un *receptive field* de $(5, 5)$.

Vemos en la figura (13b) que las predicciones son mucho más precisas que en el caso anterior, pese a que todavía existe una variabilidad importante en la determinación. Se trata de una mejora significativa frente a las redes *fully connected*. La capacidad predictiva de las redes convolucionales se explica por su arquitectura, puesto que la longitud de correlación es una medida

del tamaño medio de las zonas ordenadas. Una red que basa sus detecciones en la búsqueda de propiedades en zonas localizadas de la red (demarcadas por el *receptive field*) puede obtener de forma relativamente sencilla esta información. Es evidente entonces que un mayor tamaño del *receptive field* puede dar lugar a mejores predicciones, al ser capaz de abarcar zonas ordenadas más grandes, y delimitar correctamente las más pequeñas. Esto explica por qué los resultados mejoran ligeramente al pasar de un *receptive field* de (3,3) a (5,5), y justifica el uso de este último para nuestra red. Los resultados no mejoran más allá de este tamaño, precisamente porque hay muy pocas configuraciones con una longitud de correlación por encima de 5, y el resto de configuraciones se pueden caracterizar bien con un *receptive field* menor.

Por último, queda por estudiar el comportamiento de una red convolucional en la que el *output* no se encuentre limitado. Para ello, cambiamos la neurona *sigmoid* a la salida por una neurona $ReLU$, cuyo *output* puede tomar cualquier valor positivo. En este caso no podemos usar la función *cross-entropy* para calcular el coste, y recurriremos a una función coste de mínimos cuadrados (ver anexo 1). Como vemos en la figura (14a), los resultados son muy similares a los casos anteriores, situando correctamente el valor máximo y ajustándose en general para todas las β. De nuevo, se observa que la red tiene algunos problemas para consensuar las predicciones en la fase ordenada según la orientación del spin. Vemos en la figura (14) que las predicciones se ajustan razonablemente al valor numérico, de forma similar al caso anterior.



(a) (b)

Figura 14: Predicciones para una red de Ising ferromagnética con $L = 40$ haciendo uso de una red neuronal convolucional con una neurona $ReLU$ a la salida. (a) Predicciones y resultados numéricos para el valor medio de la longitud de correlación en función de β. El máximo se sitúa de nuevo en $\beta_c = 0{,}422 \pm 0{,}04$. (b) En verde, predicciones para la longitud de correlación frente al valor obtenido de forma numérica. La línea dorada marca la diagonal, el valor que deberían tomar las predicciones.

En definitiva, las redes convolucionales se muestran claramente superiores a las *fully connected* en la predicción de observables, hecho que se explica por la incapacidad de las últimas para determinar correctamente la topología local de las redes de Ising. En cuanto a las redes convolucionales, el comportamiento resulta muy similar independientemente de la neurona que coloquemos a la salida. Las redes con *output* limitado tienen el problema de que todas sus predicciones deben ser reescaladas adecuadamente, y además tienen un límite superior a la longitud de correlación que son capaces de predecir dado por el valor máximo con el que han sido entre-

nadas. De aquí se concluye que las redes convolucionales con un *output* ilimitado son las mejores en la predicción de la longitud de correlación. No obstante, incluso mostrando malos resultados en las predicciones para configuraciones concretas, las tres redes logran situar correctamente los máximos en función de β, y su magnitud.

7. Conclusiones

En el presente trabajo hemos comprobado que las redes neuronales *fully connected* y las convolucionales son capaces de aprender y predecir distintas características del modelo de Ising bidimensional, permitiéndonos extraer de estas predicciones información acerca de la existencia y particularidades de su transición de fase.

En primer lugar, hemos logrado entrenar redes en la clasificación de configuraciones según su pertenencia a una fase ordenada y desordenada, lo que nos ha permitido establecer una temperatura crítica de frontera entre ambas fases. Esta frontera ha resultado dependiente del tamaño de la red de Ising y del tipo de red neuronal utilizada: las redes *fully connected* han mostrado su capacidad de cálculo de observables generales como la magnetización, y las convolucionales, su comprensión del orden local. En segundo lugar, hemos logrado entrenar redes en la predicción de la longitud de correlación, obteniendo buenas estimaciones incluso para la configuraciones con longitudes más grandes. En este caso las redes convolucionales han obtenido resultados muy satisfactorios, pero ha quedado patente la limitación que conlleva el desconocimiento de la topología del sistema en las redes *fully connected*, que da lugar a un buen comportamiento en promedio pero algo pobre en los resultados para configuraciones individuales.

En definitiva, las redes neuronales se han mostrado capaces de extraer información útil acerca del modelo de Ising y aplicarla en sus predicciones. Incluso con la simpleza del modelo, hemos podido comprobar las limitaciones y características más importantes de cada tipo de red neuronal y algunos de sus hiperparámetros, recorriendo los primeros pasos en el camino de la implantación de la Inteligencia Artificial como herramienta básica en el estudio de la materia condensada.

Referencias

[1] Mehta, P., Bukov, M., Wang, C. H., Day, A. G., Richardson, C., Fisher, C. K., & Schwab, D. J. (2019). *A high-bias, low-variance introduction to machine learning for physicists*. Physics reports, 810, 1-124.

[2] Nielsen, M. A. (2015). *Neural networks and deep learning*. San Francisco, CA: Determination press.

[3] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). *Learning representations by back-propagating errors.* Nature, 323(6088), 533-536.

[4] Onsager, L. (1944). *Crystal Statistics. I. A two-dimensional model with an oder-disorder transition.* Phys. Rev., 65, 117-149.

[5] McCoy, B. M., & Wu, T. T. (2014). *The two-dimensional Ising model*. Courier Corporation.

[6] Carrasquilla, J., & Melko, R. G. (2017). *Machine learning phases of matter.* Nature Physics, 13(5), 431-434.