Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Matemáticas

Métodos numéricos para la valoración de derivados financieros

Autor: Pablo Benito Morate

Tutores: Víctor Gatón Bustillo

Beatriz Gómez Martín

2023/2024

A mi abuelo Raquel

Índice general

Resumen

La valoración de derivados financieros es un problema de gran relevancia en la sociedad actual. En este trabajo, se presentarán las herramientas matemáticas fundamentales, basadas en la Teoría de la Probabilidad y los Procesos Estocásticos, necesarias para modelar un mercado financiero. Se explicarán los modelos de Black-Scholes y Heston. Posteriormente, se describirán los modelos de redes neuronales y se expondrán los resultados obtenidos al aplicarlos para aproximar el precio de una opción Call Europea viendo su potencialidad para el problema de valoración en tiempo real.

Abstract: The valuation of financial derivatives is a problem of great relevance in today's society. In this paper, the fundamental mathematical tools, based on Probability Theory and Stochastic Processes, needed to model a financial market will be presented. The Black-Scholes and Heston models will be explained. Subsequently, the neural network models will be described and the results obtained by applying them to approximate the price of a European Call option will be presented, showing their potential for the real time trade valuation problem.

Agradecimientos

En estas breves líneas deseo expresar mi más profunda gratitud a quienes me han ayudado a sacar adelante este trabajo.

A mi tutor Víctor Gatón por introducirme en el fascinante mundo de las Finanzas Cuantitativas, por el maravilloso trato que he recibido de su parte y por toda la paciencia y consideraciones que ha tenido conmigo durante la realización de esta memoria.

A mi tutora Beatriz Gómez por su inestimable ayuda con la parte de Redes y Programación.

A mi Familia, porque gracias a todo su esfuerzo y cariño soy quien soy y en especial a mi Tío Juanjo por transmitirme desde pequeño su pasión por la ciencia.

A mis amigos, por habernos peleado juntos con tantos problemas y por ser parte de los recuerdos inolvidables de estos años.

Introducción

Los derivados financieros son unos instrumentos cuyo valor se determina a partir del precio de otro activo del mercado, conocido como subyacente. Aunque diversas formas de derivados financieros han sido objeto de negociación desde la Antigüedad, no fue hasta 1697 que aparecen los primeros registros de un mercado de futuros y derivados organizado y regulado con la Bolsa de Arroz de Dojima, en Osaka, Japón.

Este mercado sentó las bases para el desarrollo posterior de mercados financieros más sofisticados, introduciendo la idea de contratos estandarizados y regulados y, utilizando los derivados financieros de manera pionera en el comercio del arroz, un producto vital para la economía japonesa. A través del uso de contratos derivados, se lograba mitigar la incertidumbre causada por factores impredecibles como el clima, las plagas y las variaciones en la oferta y demanda, lo que contribuía a estabilizar los precios para los consumidores finales.

Antes del año 1900, debido a la ausencia de modelos matemáticos avanzados, el precio de los derivados solía determinarse de manera intuitiva mediante estimaciones imprecisas que realizaban los comerciantes basadas en su experiencia previa sobre la oferta y la demanda futuras. Fue en ese año cuando Louis Bachelier (1870-1946), bajo la supervisión de Henri Poincaré, introdujo en su tesis doctoral titulada *Thèorie de la Speculation* los fundamentos matemáticos para poder valorar opciones de manera sistemática y rigurosa.

Bachelier no obtuvo un reconocimiento significativo en vida, sin embargo, sus trabajos fueron adelantados a su tiempo ya que fue pionero en la aplicación de la Teoría de la Probabilidad y los Procesos Estocásticos al estudio de los mercados financieros. No obstante, la teoría de Bachelier no se ajusta de manera adecuada a la realidad del Mercado.

En 1973, Fischer Black y Myron Scholes en [7], y de manera independiente Robert Merton en [26], desarrollaron un marco teórico para la valoración de opciones asumiendo que los activos del mercado seguían un movimiento browniano

geométrico, hecho que se ajustaba mejor a las observaciones empíricas del mercado.

Estas innovaciones teóricas coincidieron con la inauguración del CBOE (Chicago Board Options Exchange), la primera bolsa de opciones del mundo, establecida en 1973 en Chicago, Estados Unidos. Esta circunstancia condujo a la rápida adopción del modelo de Black-Scholes como un método efectivo de valoración. Como resultado de sus contribuciones, Scholes y Merton fueron galardonados en 1997 con el Premio Nobel de Economía, aunque Black no pudo recibirlo debido a que los premios Nobel no se otorgan de forma póstuma, y él falleció dos años antes.

El modelo de Black-Scholes presenta una serie de limitaciones que hacen que sus valoraciones no encajen del todo con la realidad observada, por ejemplo, en periodos de alta volatilidad en el mercado. Esto ha impulsado el desarrollo de nuevos modelos mas complejos y realistas pero que requieren de métodos numéricas para su resolución.

En la actualidad, los métodos numéricos utilizados abarcan una amplia gama, que incluye desde la resolución numérica de ecuaciones diferenciales en derivadas parciales hasta técnicas de simulación estadística. En esta memoria, se examinará el potencial de uso de las redes neuronales como herramienta para la valoración de productos financieros en tiempo real, problema conocido como *real-time trade*. Aunque las redes fueron planteadas por primera vez en la década de 1950, su importancia comenzó a aumentar significativamente hacia mediados de la década de 1980. Esto se debió al desarrollo del algoritmo de *Backpropagation* y los avances en la capacidad computacional, lo cual permitió su aplicación efectiva a una amplia gama de problemas.

Capítulo 1

Mercados y Derivados Financieros

En este capítulo se describirán los principales elementos, de un mercado financiero, haciendo énfasis en los derivados. También se introducirá el problema de valoración de un derivado y las herramientas matemáticas e hipótesis necesarias para la modelización del mismo.

1.1. Definiciones y Primeras Propiedades

Un *Mercado* es un lugar donde se encuentran individuos con necesidad de un bien junto con otros individuos que tienen un excedente de dicho bien.

Un *Mercado Financiero* es, por tanto, un mercado en el que se negocian e intercambian instrumentos financieros.

En un mercado, el dinero existe para facilitar las transacciones y establecer los precios de los productos que se negocian en él. Por ejemplo, a las diferentes monedas que emiten los Estados, se las denominan *divisas*. Las divisas conforman un instrumento financiero, ya que pueden ser intercambiadas entre sí mediante tipos de cambio que fluctúan de manera constante en función de los factores económicos y socio-políticos de cada nación.

Las materias primas o *commodities* también se negocian en un mercado financiero, debido a que representan una forma de valor con uniformidad. Entre las commodities más negociadas están los productos agrícolas como el trigo y el café, los metales como el oro y el cobre, y los recursos energéticos como el petróleo y el gas natural. Los precios de las materias primas son en general, impredecibles debido a la influencia de factores geopolíticos, económicos y ambientales.

Otro de los motivos por el que las personas e instituciones suelen acudir a

los mercados financieros es para obtener liquidez. Para ello se pueden emplear los *bonos*, que son instrumentos financieros que, a cambio de un préstamo de capital al emisor del bono, ofrecen en una fecha futura la devolución del capital prestado junto unos intereses fijados de antemano. Como en general, el retorno futuro de un bono es conocido, se dice que es un *Activo Libre de Riesgo*.

El retorno final de un bono esta completamente determinado al principio de la inversión. Por comodidad se supondrá que todo activo libre de riesgo se comporta igual que un bono, en el que si se invierte un capital C_0 en el instante t, este devuelve $C_0 e^{r(T-t)}$ en el instante T con $t < T$, donde r, es lo que se conoce como el tipo de interés libre riesgo.

Otra forma de obtener liquidez para una compañía es, por ejemplo, la emisión de *acciones*. Al emitir acciones en el mercado financiero, la compañía está ofreciendo una participación de propiedad a los inversores a cambio de capital. Esta participación concede además a los accionistas la posibilidad de participar en la toma de decisiones de la compañía mediante el derecho a voto en junta de accionistas, así como el derecho a recibir dividendos, que son pagos que la compañía efectúa como compensación por su inversión a los accionistas con motivo de distribuir las ganancias obtenidas en un periodo determinado. El precio de una acción se puede ver afectado por el desempeño de la empresa o por el estado general de la economía. Como el futuro no se conoce *a priori*, el valor de cada participación varía de forma a veces impredecible, por lo que la inversión en acciones supone una inversión con riesgo para el accionista.

La gestión del riego es, de hecho, uno de los principales motivos por el que los los individuos acuden a los mercados financieros. Para desarrollar estrategias de cobertura ante el riesgo (*Hedging*), existen unos instrumentos financieros denominados *Derivados*, que consisten en contratos cuyo valor se determina en función del precio de otro activo del mercado, denominado *Activo Subyacente* del derivado. Los activos subyacentes pueden de ser de diversos tipos, incluyendo acciones, bonos, commodities, divisas, etc.

Los derivados permiten a una de las partes del contrato transferir un riesgo a cambio del pago de una prima, a otra parte que este dispuesta a asumirlo a cambio del pago de esa prima.

El derivado financiero más simple que existe es el *Futuro*. Este contrato obliga a una de las partes a comprar, y a la otra a vender, una cantidad específica del activo subyacente a un precio previamente acordado en una fecha futura específica, conocida como vencimiento o *maturity*. Si este tipo de derivado se negocia fuera de un mercado financiero organizado, como una bolsa de valores, se denomina contrato *forward*.

Las *Opciones*, son otro tipo de derivado que, a cambio del pago de una prima, ofrecen al comprador de la opción el derecho, pero no la obligación, de comprar (Opción de Compra o *Call*) o vender (Opción de Venta o *Put*) una cantidad específica del activo subyacente por el precio de ejercicio o *Strike* que se haya fijado.

Dentro de las opciones, existen diferentes clases de contratos en función de las condiciones que ofrecen. Los tipos de opciones mas negociados son:

- **Opción Europea:** Este contrato otorga al comprador de la opción el derecho de comprar (Call) o vender (Put) el activo subyacente, por el precio de ejercicio acordado, únicamente en una fecha futura T denominada vencimiento. El titular de la opción, únicamente ejerce su derecho (i.e. *ejecuta* la opción) si el valor de la opción en el vencimiento (su *payoff*) le es favorable.

  Si se considera una Call Europea de strike K, entonces el *payoff* de dicha opción viene dado por

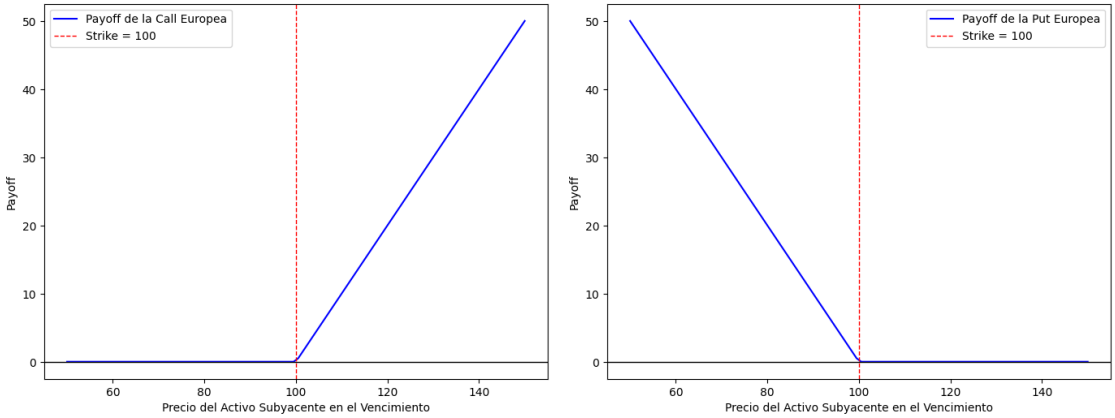
  $$\text{máx}(S(T) - K, 0)\,, \tag{1.1}$$

  donde $S(T)$ es el valor del activo subyacente en el vencimiento.

  Si se considera una Put Europea de strike K, su *payoff* viene determinado por la expresión

  $$\text{máx}(K - S(T), 0)\,. \tag{1.2}$$

  En la Figura 1.1 se puede ver representado, a la derecha, el diagrama del *payoff* de una opción de compra Europea con strike 100 en función del valor del activo subyacente en el momento del vencimiento. A la izquierda de la figura, se muestra lo mismo pero para la Put Europea.



(a) *Call* Europea (b) *Put* Europea

Figura 1.1: *Payoffs* de una opción Europea de strike $K = 100$ en función del precio del activo subyacente en el vencimiento.

- **Opción Americana:** La opción americana, es similar a la Europea con la diferencia de que la opción americana se puede ejercer en cualquier instante anterior o igual al vencimiento, y no únicamente en la fecha de vencimiento como en el caso de la opción Europea.

- **Opción Bermuda:** La opción bermuda combina características de las opciones europeas y americanas. Permite comprar o vender una cantidad específica del activo subyacente por el precio de ejercicio que se haya fijado, pero únicamente en una serie de fechas de ejercicio establecidas de antemano.

- **Opción Asiáticas:** Las opciones que se han detallado hasta ahora, cuando se ejecutan, se liquidan con el valor exacto del activo subyacente en la fecha de vencimiento. Sin embargo, las opciones asiáticas dependen del valor medio que haya tenido el activo subyacente durante la vigencia del contrato.

  Este tipo de contratos permite a los inversores apostar por la tendencia que puede tener el precio de cierto activo, sin preocuparse por las posibles fluctuaciones extremas que pueda presentar en fechas cercanas al vencimiento.

- **Opción Barrera:** El *payoff* de las opciones barrera se realiza si y solo si, el precio del activo subyacente alcanza (o no ha alcanzado), un valor previamente acordado, llamado valor barrera, antes de la fecha de vencimiento de la opción.

  Las opciones barrera se clasifican principalmente en dos tipos: *knock-in*, si el *payoff* se cobra (en el vencimiento) si el precio del subyacente alzanza la

barrera a lo largo de la trayectoria y, *knock-out* si el *payoff* deja de cobrarse si el precio del subyacente alcanza la barrera.

En un mercado financiero, se permite que los individuos tomen posiciones largas o cortas en los diferentes activos financieros. Una posición larga supone la compra de un activo con la expectativa de que su valor aumente en el futuro, lo que permite obtener ganancias al venderlo a un precio más alto. Por otro lado, una posición corta implica vender un activo que el inversor no posee realmente en el momento de la venta, con la intención de comprarlo más tarde a un precio más bajo para obtener ganancias. Esta práctica, conocida como "vender en corto" o (*short-selling*), se permite en los mercados financieros y posibilita a los inversores beneficiarse de caídas en el precio de los activos (por ejemplo mediante futuros).

Los inversores en el mercado financiero no están restringidos a mantener una posición en un solo activo. Tienen la capacidad de desarrollar estrategias diversificadas mediante la compra y venta de una variedad de activos de diferentes clases. Esta colección de inversiones variadas se conoce como *Portfolio* (o portfolio de inversión).

1.1.1. La Hipótesis de No Arbitraje

Aunque existen varias definiciones similares, esencialmente, una oportunidad de arbitraje es una estrategia financiera que, en un escenario con riesgo (i.e del que se desconoce su evolución futura), permite obtener ganancias sin necesidad de un aporte inicial de capital.

La hipótesis de No Arbitraje asume la inexistencia de oportunidades de arbitraje en un mercado financiero. El fundamento de esta hipótesis, consiste en el hecho de que si existiera una diferencia de precios entre dos activos que supuestamente deberían de ser iguales, esta no podría existir de manera sostenida en el mercado, ya que los inversores del mercado explotarían rápidamente esta oportunidad, ajustando los precios en el proceso y haciendo desaparecer la oportunidad de arbitraje.

En el contexto de la valoración de derivados financieros es fundamental asumir la hipótesis de no arbitraje. Esta suposición permite, a partir de principios económicos inmediatos, obtener una serie de resultados que tienen que cumplir necesariamente los precios de algunos derivados para evitar incurrir en oportunidades de arbitraje.

Para los siguientes resultados, se fija el convenio de que los flujos de capital que representan ingresos o ganancias son positivos y que los flujos de capital que representan egresos o pérdidas son negativos.

Proposición 1.1. *El tipo de interés libre riesgo es único.*

Demostración. Supongamos que existen dos activos libres de riesgo con tipos de intéres r_1 y r_2 respectivamente. Sin pérdida de generalidad, podemos suponer que $r_1 < r_2$.

Es posible conformar un portfolio vendiendo en corto una cantidad A del bono con el menor tipo de interés (esta operación es equivalente a pedir prestadas A unidades monetarias a un banco) y reinvirtiéndolo en la compra del bono con el tipo de interés mayor.

	Valor en t	Valor en T
Venta Bono r_1	A	$-Ae^{r_1(T-t)}$
Compra Bono r_2	$-A$	$Ae^{r_2(T-t)}$
Neto:	0	$A(e^{r_2(T-t)} - e^{r_1(T-t)}) > 0$

Tabla 1.1: Ejemplo de oportunidad de arbitraje con dos tipos de interés libres de riesgo.

Entonces, como se observa en la Tabla 1.1, sin necesidad de un aporte inicial de capital y sin ninguna incertidumbre, se obtiene una cantidad estrictamente positiva de dinero, lo que es absurdo si se asume que en el mercado no pueden darse oportunidades de arbitraje. □

Para el siguiente resultado, se denota por $C(t)$ al precio que tiene en el instante t una opción Call Europea, negociada sobre el activo subyacente S, de vencimiento T y strike K. De la misma manera, se denota por $P(t)$ al precio que tiene en el instante t una opción Put Europea, negociada con el mismo strike K, vencimiento T y sobre el mismo activo subyacente S que la Call anterior.

Entonces, con la notación que se ha introducido, se tiene el siguiente resultado.

Proposición 1.2 (Paridad *Put-Call*)**.** *Para todo $t < T$ se tiene que:*

$$P(t) = C(t) + Ke^{-r(T-t)} - S(t)\,. \tag{1.3}$$

Demostración. Se razona por reducción al absurdo.

Supongamos que $-P(t) + C(t) + Ke^{-r(T-t)} - S(t) = A > 0$.

Si en el instante t se crea un portfolio mediante la compra de una Put Europea, la venta de una Call Europea, la compra de una unidad del activo S y la venta de

un bono por valor de $Ke^{-r(T-t)}$, se obtiene un portfolio con el flujo de caja que aparece representado en la Tabla 1.2

	Valor en t	Valor en T
Compra *Put*	$-P(t)$	$\text{máx}(K - S(T), 0)$
Venta *Call*	$C(t)$	$-\text{máx}(S(T) - K, 0)$
Venta Bono	$Ke^{-r(T-t)}$	$-K$
Compra Activo	$-S(t)$	$S(T)$
Neto:	$A > 0$	0

Tabla 1.2: Portfolio del Caso $A > 0$.

En la Tabla 1.2 se puede observar también que el valor del portfolio en el vencimiento T es 0, ya que siempre se cumple que

$$\text{máx}(K - S(T), 0) - \text{máx}(S(T) - K, 0) = K - S(T)\,, \tag{1.4}$$

independiente del valor desconocido $S(T)$.

Entonces, si como se muestra en la Tabla 1.3, en t se vende el portfolio por su valor A y se reinvierte esa cantidad en la compra de un bono,

	Valor en t	Valor en T
Venta Portfolio	A	0
Compra Bono	$-A$	$Ae^{-r(T-t)}$
Neto:	0	$Ae^{-r(T-t)} > 0$

Tabla 1.3: Oportunidad de arbitraje si no se da la paridad Put-Call.

se demuestra que, sin necesidad de un aporte inicial de capital y sin ninguna incertidumbre, se obtiene una cantidad estrictamente positiva de dinero, lo que incumple la hipótesis de no arbitraje.

El caso $A < 0$ se razona de manera análoga al caso anterior pero conformando esta vez el porfolio mediante mediante la venta de una Put, la compra de una Call, la venta de una unidad del activo $S(t)$ y la compra de un bono por $Ke^{-r(T-t)}$.

□

En [14] se presenta un compendio completo de resultados obtenidos únicamente a partir de la aplicación de la hipótesis de No Arbitraje. No obstante, estos resultados por sí solos no son suficientes para la valoración de derivados, ya que es un problema que, para su tratamiento, requiere de un aparato matemático más complejo

1.2. Modelización Matemática

En estas sección se introducirán las herramientas matemáticas necesarias abordar el problema de determinar el precio de una opción.

1.2.1. Procesos Estocásticos y Cálculo de Itô.

Para estudiar la valoración de derivados financieros en tiempo continuo, se modelizarán los precios de los activos del mercado como procesos estocásticos continuos. Para ello, se recuerdan las siguientes definiciones de Teoría de la Probabilidad y del Cálculo Estocástico (un desarrollo detallado de los resultados aquí presentados se puede encontrar en [5] y [6]).

Definición 1.3. Sea Ω un conjunto arbitrario. Un subconjunto no vacío $\mathcal{F} \subset \mathcal{P}(\Omega)$ se dice que es una *σ-álgebra* de Ω si satisface:

1. $\Omega \in \mathcal{F}$.
2. Si $A \in \mathcal{F}$, entonces $\Omega \setminus A \in \mathcal{F}$. (Cerrada para complementarios.)
3. Si $\{A_k\}_{k=1}^{\infty} \subset \mathcal{F}$, entonces $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$. (Cerrada para uniones numerables.)

Definición 1.4. Sea $(\Omega, \mathcal{F})$ un espacio medible. Se dice que la aplicación $\mathbb{P} : \Omega \longrightarrow \mathbb{R}$ es una *medida de probabilidad* sobre $(\Omega, \mathcal{F})$ si satisface:

1. $0 \leq \mathbb{P}(A) \leq 1 \ \forall \ A \in \mathcal{F}$.
2. $\mathbb{P}(\varnothing) = 0$ y $\mathbb{P}(\Omega) = 1$
3. Para toda sucesión $\{A_k\}_{k=1}^{\infty} \subset \mathcal{F}$ de conjuntos disjuntos dos a dos se tiene que:

$$\mathbb{P}\Big(\bigcup_{k=1}^{\infty} A_k\Big) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)\,.$$

A la terna $(\Omega, \mathcal{F}, \mathbb{P})$ se la denomina *espacio de probabilidad.*

Definición 1.5. Sea $\mathcal{E} \subset \Omega$. Se define la *$\sigma$-álgebra generada por $\mathcal{E}$* en Ω, que denotaremos por $\sigma(\mathcal{E})$, a la intersección de todas las σ-álgebras de Ω que contienen a $\mathcal{E}$.

Se llama σ-álgebra de Borel de $\mathbb{R}^n$ a la σ-álgebra generada por la topología usual de $\mathbb{R}^n$. Se denota por $\mathcal{B}(\mathbb{R}^n)$.

Definición 1.6. Sean $\mathcal{F} \subset \mathcal{P}(\Omega)$ una σ-álgebra de Ω y $A \subset \Omega$ un conjunto no vacío. La σ-álgebra de A, $\mathcal{F}_A = \{F \cap A : F \in \mathcal{F}\}$ recibe el nombre de *σ-álgebra inducida en A por $\mathcal{F}$*.

Para $T > 0$ se denota por $\mathcal{B}([0,T])$ a la σ-álgebra inducida en $[0,T]$ por la σ-álgebra de Borel de $\mathbb{R}$.

Definición 1.7. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad y $\mathbb{T}$ un conjunto. Un *Proceso Estocástico* X es una colección de variables aleatorias $\{X(t)\}_{t\in\mathbb{T}}$ definidas en $(\Omega, \mathcal{F}, \mathbb{P})$ y con valores en $\mathbb{R}^n$.

A $\mathbb{P}$ se la suele llamar la medida física de probabilidad.

Nota. Generalmente $\mathbb{T}$ se suele interpretar como un conjunto de índices que especifica los momentos en el tiempo en los que se observa el proceso. $\mathbb{T}$ puede ser finito o $\mathbb{N}$ (en tal caso se denomina proceso *discreto*), pero también puede ser $[0,T]$ o $[0,\infty)$, en cuyo caso, se denomina proceso *en tiempo continuo*.

Observación 1. Un proceso estocástico se puede interpretar como una aplicación

$$\begin{array}{rccc} X\colon & \Omega\times\mathbb{T} & \longrightarrow & \mathbb{R}^n \\ & (\omega,t) & \longmapsto & X(t,\omega) \end{array}.$$

Para cada $\omega \in \Omega$, a la función $X(\omega,\cdot)\colon \mathbb{T} \longrightarrow \mathbb{R}^n$ se la denomina *realización o trayectoria* del proceso.

Definición 1.8. Se dice que una variable aleatoria real X sigue una distribución normal de media μ y desviación típica σ si tiene función de densidad dada por:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)}{2\sigma^2}\Big), \qquad x \in \mathbb{R}\,. \tag{1.5}$$

En este caso, se denota $X \sim \mathcal{N}(\mu,\sigma)$.

Definición 1.9. Sea X una variable aleatoria continua con función de densidad f_X. Se define la función característica de la variable X por

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \int_{\mathbb{R}} e^{itX} f_X(t)dt\,, \qquad \forall t \in \mathbb{R}\,. \tag{1.6}$$

El siguiente resultado, que se empleará posteriormente, se puede encontrar en [17].

Teorema 1.10 (Teorema de Inversión de Gil-Peláez)**.** *Sea X una variable aleatoria con función característica φ_X.*

Entonces se tiene que si F_X es la función de distribución de X, esta viene dada por

$$F_X(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^\infty \mathrm{Im}\Big[\frac{e^{-itx}\varphi_X(t)}{t}\Big]dt\,, \qquad x \in \mathbb{R}\,. \tag{1.7}$$

Definición 1.11. Un proceso estocástico continuo W se denomina *Movimiento Browniano* o *Proceso de Wiener* si verifica que:

1. $W(0) = 0$
2. Para todo $s < t$ se tiene que $W(t) - W(s) \sim \mathcal{N}(0, \sqrt{t-s})$.
3. W tiene *incrementos independientes*. Es decir, si $r < s \leq t < u$ entonces las variables aleatorias $W(u) - W(t)$ y $W(s) - W(r)$ son independientes.
4. La trayectoria de W es continua.

El objetivo es describir la dinámica de un proceso que se aproxima localmente por un término determinista y otro que sigue un proceso de Wiener, cuya dinámica viene dada por:

$$\begin{cases} dX(t) = \mu(t, X(t))dt + \sigma(t, X(t))dW(t) \\ X(0) = a \end{cases}\,, \tag{1.8}$$

donde a $\mu(t, X(t))$ se le denomina *drift* y $\sigma(t, X(t))$ es un término de difusión que modeliza la amplitud que la componente gaussiana tiene en cada instante.

Dado un proceso estocástico X, es importante precisar el concepto de *la información generada por X*. Para ello, se introduce el concepto de filtración.

Definición 1.12. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad y sea I un conjunto de índices totalmente ordenado. Una *filtración* en $(\Omega, \mathcal{F}, \mathbb{P})$ es una colección de σ-álgebras de $\mathcal{F}$, $\mathbb{F} = \{\mathcal{F}_i\}_{i\in I}$ tales que

$$\mathcal{F}_k \subseteq \mathcal{F}_j \text{ para cada } k \leq j\,.$$

Definición 1.13. Sea $\{X(t)\}_{t\geq 0}$ un proceso estocástico definido en el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$. Se define la *$\sigma$-álgebra generada por X en el intervalo* $[0, t]$ como

$$\mathcal{F}_t^X = \sigma\Big(\{X(s)\colon 0 \leq s \leq t\}\Big)\,,$$

es decir, la menor σ-álgebra de Ω que contiene a $\{X(s)\colon 0 \leq s \leq t\}$.

Si $A \in \mathcal{F}_t^X$, se dirá que A es $\mathcal{F}_t^X$-medible.

Observación 2. Para un proceso estocástico $\{X(t)\}_{t\geq 0}$ definido en $(\Omega, \mathcal{F}, \mathbb{P})$, la colección $\{\mathcal{F}_t^X\}_{t\geq 0}$ es una filtración de $(\Omega, \mathcal{F}, \mathbb{P})$.

Definición 1.14. Sea $\{X(t)\}_{t\geq 0}$ un proceso estocástico definido en el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ y sea $\{\mathcal{F}_t^X\}_{t\geq 0}$ una filtración de $(\Omega, \mathcal{F}, \mathbb{P})$.

Diremos que el proceso $\{Y(t)\}_{t\geq 0}$ está *adaptado a la filtración* $\{\mathcal{F}_t^X\}_{t\geq 0}$ si para todo $t \geq 0$ se tiene que

$$Y(t) \in \mathcal{F}_t^X$$

Definición 1.15. Sea $\{X(t)\}_{t\geq 0}$ un proceso estocástico definido en el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ y sea $\mathbb{F} = \{\mathcal{F}_t\}_{t\geq 0}$ una filtración en $(\Omega, \mathcal{F}, \mathbb{P})$. Se dice que el proceso estocástico X es una $\{\mathcal{F}_t\}_{t\geq 0}$-*martingala* si se satisface que:

1. X está adaptado a la filtración $\{\mathcal{F}_t\}_{t\geq 0}$.
2. $\mathbb{E}[|X(t)|] < \infty$ para todo $t \geq 0$.
3. $\mathbb{E}[X(t)|\mathcal{F}_s] = X(s)$ para todo s, t con $s \leq t$.

Es natural interpretar la expresión (1.8) como la ecuación integral

$$X(t) = a + \int_0^t \mu(s, X(s))ds + \int_0^t \sigma(s, X(s))dW(s)\,. \tag{1.9}$$

No obstante, el tercer término de la derecha no se puede interpretar como una integral de Riemann-Stieltjes en el sentido usual, ya que las trayectorias del proceso de Wiener son localmente no acotadas.

Como se expone en [22], en la década de 1940, el japonés Kiyoshi Itô desarrolló una teoría de integración estocástica para abordar el problema anterior, la cual se conoce como el Cálculo de Itô.

Una construcción completa de la integral estocástica de Itô para un proceso de cuadrado integrable arbitrario se puede encontrar en [8], aunque esquemáticamente se procede de la siguiente manera, como se presenta en [6].

Definición 1.16. Sea $\{W(t)\}_{t\geq 0}$ un proceso de Wiener. Un proceso estocástico g se dice que pertenece a $L^2([a, b])$, o que es de cuadrado integrable en $[a, b]$, si verifica que:

1. $\int_a^b \mathbb{E}[g^2(s)]ds < \infty$.
2. g está adaptado a la filtración $\{\mathcal{F}_t^W\}_{t\geq 0}$.

Definición 1.17. Se dice que un proceso estocástico $g \in L^2([a,b])$ es *simple* si existe una sucesión determinista de puntos $a = t_0 < t_1 < \cdots < t_n = b$ tales que

$$g(s) = g(t_k) \text{ para todo } s \in [t_k, t_{k+1}) .$$

Definición 1.18. Para un proceso estocástico simple $g \in L^2([a,b])$ se define la *integral estocástica de Itô* como

$$\int_a^b g(s)dW(s) = \sum_{k=0}^{n-1} g(t_k)[W(t_{k+1}) - W(t_k)] .$$

Entonces, en primer lugar, se demuestra que para un proceso $g \in L^2([a,b])$ existe una sucesión $\{g_n\}_{n=1}^{\infty}$ de procesos simples de $L^2([a,b])$ tales que

$$\int_a^b \mathbb{E}\Big[(g_n(s) - g(s))^2\Big] ds \longrightarrow 0 .$$

Así mismo, se prueba que la variable aleatoria $Z_n = \int_a^b g_n(s)dW(s)$ esta bien definida y que existe una variable aleatoria Z tal que $Z_n \xrightarrow[n\to\infty]{} Z$ en L^2.

Por último, se define la *Integral Estocástica de Itô* por

$$\int_a^b g(s)dW(s) = \lim_{n\to\infty} \int_a^b g_n(s)dW(s) .$$

Definición 1.19. Sea $\{X(t)\}_{t\geq 0}$ proceso estocástico definido en el espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$.

Se dice que X es un *Proceso de Itô* si existen un proceso de Wiener $\{W(t)\}_{t\geq 0}$, un proceso estocástico K adaptado a la filtración $\{\mathcal{F}_t^W\}_{t\geq 0}$ tal que $\int_0^t |K(s)|ds < \infty$ y un proceso $H \in L^2([0,t])$ para todo $t \geq 0$ tales que permiten que X admita la representación:

$$X(t) = X(0) + \int_0^t K(s)ds + \int_0^t H(s)dW(s) \text{ para todo } t \geq 0 , \tag{1.10}$$

o escrita en forma diferencial

$$dX(t) = K(t)dt + H(t)dW(t) .$$

De acuerdo con [14], si un proceso de Itô admite dos representaciones, entonces estas coinciden casi seguro.

Proposición 1.20. *Sea X un proceso de Itô que sigue la dinámica dada por:*

$$dX(t) = \mu(t)dt + \sigma(t)dW(t)\,,$$

Entonces, X es una martingala si, y solamente si $\mu(t) = 0$ casi siempre.

Demostración. Supongamos que X es una martingala, entonces, por definición se tiene que

$$\mathbb{E}[X(t)|\mathcal{F}_s] = X(s)\,.$$

.

Aplicando la definición de proceso de Itô y la linealidad de la esperanza condicionada se cumple que

$$\begin{aligned}\mathbb{E}[X(t)|\mathcal{F}_s] &= \mathbb{E}\Big[X(s) + \int_s^t \mu(u)du + \int_s^t \sigma(u)dW(u)\Big|\mathcal{F}_s\Big] \\ &= X(s) + \mathbb{E}\Big[\int_s^t \mu(u)du\Big|\mathcal{F}_s\Big] + \mathbb{E}\Big[\int_s^t \sigma(u)dW(u)\Big|\mathcal{F}_s\Big] \\ &= X(s)\,.\end{aligned}$$

A partir de la definición de proceso de Wiener se sigue que

$$\mathbb{E}\Big[\int_s^t \sigma(u)dW(u)\Big|\mathcal{F}_s\Big] = 0\,,$$

por lo que necesariamente se tiene que cumplir que

$$\mathbb{E}\Big[\int_s^t \mu(u)du\Big|\mathcal{F}_s\Big] = 0\,,$$

con lo cual, aplicando las propiedades de la esperanza condicionada, se comprueba que $\int_s^t \mu(u)du = 0$ c.s, y por las mismas razones, se tiene que verificar que $\mu = 0$ casi siempre.

Si se supone que $\mu = 0$ casi siempre, entonces la dinámica del proceso X viene dada por

$$dX(t) = \sigma(t)dW(t)\,.$$

Entonces

$$\mathbb{E}[X(t)|\mathcal{F}_s] = \mathbb{E}\Big[X(s) + \int_s^t \sigma(u)dW(u)\Big|\mathcal{F}_s\Big] = X(s)\,,$$

por lo que X es una martingala.

□

Proposición 1.21 (Lema de Itô)**.** *Sea X un proceso de Itô que sigue la dinámica dada por:*

$$dX(t) = \mu(t)dt + \sigma(t)dW(t)\,, \tag{1.11}$$

y sea $f\colon (0,\infty)\times\mathbb{R} \longrightarrow \mathbb{R}$ de clase $\mathcal{C}^{1,2}$. Entonces, $f(t,X(t))$ verifica que:

$$df(t,X(t)) = \Big(\frac{\partial f}{\partial t} + \mu\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 f}{\partial x^2}\Big)dt + \sigma\frac{\partial f}{\partial x}dW(t)\,. \tag{1.12}$$

Demostración. Para no recargar la memoria, no se realizará una demostración exhaustiva (que se puede encontrar en [10] y en [31]) sino que se presentará una aproximación de la misma que aparece en [6].

En virtud de la fórmula de Taylor para funciones de varias variables (ver [15]), para cada $(t_0,x_0)\in\mathbb{R}^2$ y $r>0$ se tiene que:

$$\begin{aligned} f(t,x) = {} & f(t_0,x_0) + \frac{\partial f}{\partial t}(t,x)(t-t_0) + \frac{\partial f}{\partial x}(t,x)(x-x_0) \\ & + \frac{1}{2}\Big(\frac{\partial^2 f}{\partial t^2}(t,x)(t-t_0)^2 + \frac{\partial^2 f}{\partial x^2}(t,x)(x-x_0)^2 + 2\frac{\partial^2 f}{\partial x\partial t}(t,x)(t-t_0)(x-x_0)\Big) \\ & + o(||(t-t_0,x-x_0)||^2)\,, \qquad \forall (t,x)\in B((t_0,x_0),r)\,. \end{aligned}$$

Denotando $\Delta f = f(t,x)-f(t_0,x_0)$, $\Delta t = (t-t_0)$, $\Delta x = (x-x_0)$ y permitiendo un pequeño abuso de notación, se puede reescribir la expresión anterior como:

$$\begin{aligned} \Delta f = {} & \frac{\partial f}{\partial t}\Delta t + \frac{\partial f}{\partial x}\Delta x \\ & + \frac{1}{2}\Big(\frac{\partial^2 f}{\partial t^2}(\Delta t)^2 + \frac{\partial^2 f}{\partial x^2}(\Delta x)^2 + 2\frac{\partial^2 f}{\partial x\partial t}\Delta t\Delta x\Big) \\ & + o(||(t-t_0,x-x_0)||^2)\,. \end{aligned}$$

Haciendo tender r hacia 0, se obtiene que:

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}dx + \frac{1}{2}\Big(\frac{\partial^2 f}{\partial^2 t^2}(dt)^2 + \frac{\partial^2 f}{\partial^2 x^2}(dx)^2 + 2\frac{\partial^2 f}{\partial x\partial t}dtdx\Big)\,,$$

donde sustituyendo x por el proceso de Itô X, se llega a que

$$df = \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}dX + \frac{1}{2}\frac{\partial^2 f}{\partial^2 t^2}(dt)^2 + \frac{1}{2}\frac{\partial^2 f}{\partial^2 x^2}(dX)^2 + \frac{\partial^2 f}{\partial x\partial t}dtdX\,. \tag{1.13}$$

Por la definición de (1.11), formalmente se tiene que:

$$(dX)^2 = \mu^2(dt)^2 + 2\mu\sigma(dt)(dW) + \sigma^2(dW)^2\,. \tag{1.14}$$

Entonces, sustituyendo (1.11) y (1.14) en (1.13) y desarrollando, se tiene que:

$$\begin{aligned} df =& \frac{\partial f}{\partial t}dt + \frac{\partial f}{\partial x}(\mu dt + \sigma dW) \\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\Big(\mu^2(dt)^2 + 2\mu\sigma dt dW + \sigma^2(dW)^2\Big) \\ &+ \frac{1}{2}\frac{\partial^2 f}{\partial t^2} + \frac{\partial^2 f}{\partial x \partial t}\Big(\mu(dt)^2 + \sigma dt dW\Big)\,. \end{aligned} \tag{1.15}$$

Como se menciona en [6], los términos que contienen $(dt)^2$ y $dtdW$ son despreciables respecto a los términos que contienen dt. Entonces, probar que $(dW)^2 = dt$ es equivalente a demostrar (1.12).

Para ello, como se muestra en [6] y en [37], se considera una subdivisión del intervalo $[0, t]$ en n partes iguales, de manera que $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = t$, donde $t_k = \frac{kt}{n}$. Para cada subdivisión, se define la *variación cuadrática* del proceso de Wiener W por:

$$S_n = \sum_{k=1}^{n} \Big(W(t_k) - W(t_{k-1})\Big)^2\,. \tag{1.16}$$

Como W es un proceso de Wiener, se cumple que $W(t_k) - W(t_{k-1}) \sim \mathcal{N}\left(0, \sqrt{\frac{t}{n}}\right)$, entonces se tiene que:

$$\begin{aligned} \mathbb{E}[W(t_k) - W(t_{k-1})] &= 0\,, \\ \mathrm{Var}[W(t_k) - W(t_{k-1})] &= \frac{t}{n}\,. \end{aligned}$$

Como:

$$\mathrm{Var}[W(t_k) - W(t_{k-1})] = \mathbb{E}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big] - \mathbb{E}[W(t_k) - W(t_{k-1})]^2\,,$$

se obtiene que:

$$\mathbb{E}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big] = \frac{t}{n} \text{ para cada } k = 1\ldots, n\,. \tag{1.17}$$

Entonces, aplicando la linealidad de la esperanza se obtiene que:

$$\mathbb{E}[S_n] = \sum_{k=1}^{n} \mathbb{E}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big] = \sum_{k=1}^{n} \frac{t}{n} = t\,. \tag{1.18}$$

Para calcular la varianza de S_n, se acude al hecho de que los procesos de Wiener tienen incrementos independientes, con lo cual:

$$\mathrm{Var}[S_n] = \sum_{k=1}^{n} \mathrm{Var}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big] .$$

Puesto que:

$$\mathrm{Var}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big] = \mathbb{E}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^4\Big] - \mathbb{E}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big]^2 ,$$

y, como es conocido que el momento central de orden cuatro de una variable aleatoria normal $Z \sim \mathcal{N}(0, \sigma)$ es $\mathbb{E}[Z^4] = 3\sigma^2$, se tiene que:

$$\mathrm{Var}\Big[\Big(W(t_k) - W(t_{k-1})\Big)^2\Big] = 3\Big(\frac{t}{n}\Big)^2 - \Big(\frac{t}{n}\Big)^2 = 2\Big(\frac{t}{n}\Big)^2 \text{ para cada } k = 1, \ldots, n .$$

Con lo cual, se tiene que:

$$\mathrm{Var}[S_n] = \sum_{k=1}^{n} 2\Big(\frac{t}{n}\Big)^2 = \frac{2t^2}{n} \xrightarrow[n\to\infty]{} 0 . \tag{1.19}$$

por lo tanto, se ve que cuando $n \to \infty$, S_n tiende hacia el límite t con varianza 0, por lo tanto, en virtud de la construcción que se ha descrito antes para la integral estocástica de Itô, se puede considerar que:

$$\int_0^t (dW)^2 = t ,$$

o equivalentemente:

$$(dW)^2 = dt . \tag{1.20}$$

Así pues, despreciando en (1.15) los términos que contienen $(dt)^2$ y $dtdW$, sustituyendo $(dW)^2$ por dt y reordenando los términos, se tiene que:

$$df = \Big(\frac{\partial f}{\partial t} + \mu\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2\frac{\partial^2 f}{\partial x^2}\Big)dt + \sigma\frac{\partial f}{\partial x}dW ,$$

como se quería demostrar. □

Se puede dar una versión de este resultado en varias variables (ver [6] y [31]).

Definición 1.22. Un proceso n-dimensional de Itô es un proceso vectorial $X = (X_1, \cdots, X_n)^T$ donde cada componente X_i sigue un proceso estocástico dado por

$$dX_i(t) = \mu_i(t)dt + \sum_{j=1}^{n} \sigma_{i,j}(t)dW_j(t)\,,$$

y donde cada dW_i denota a un proceso de Wiener que tiene una correlación con el resto de procesos denotada por

$$\rho_{i,j}dt = \mathrm{Cov}(dW_i, dW_j)\,, \qquad j = 1\ldots, n\,.$$

Proposición 1.23 (Lema de Itô Multidimensional)**.** *Sea X un proceso de Itô n-dimensional y sea $f\colon (0,\infty)\times\mathbb{R}^n \longrightarrow \mathbb{R}$ de clase $\mathcal{C}^{1,2}$. Entonces, $f(t, X(t))$ verifica que:*

$$df(t, X(t)) = \frac{\partial f}{\partial t}dt + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}dX_i + \frac{1}{2}\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}dX_i dX_j \tag{1.21}$$

junto con la siguiente tabla de multiplicación

$$\begin{cases} (dt)^2 = 0, \\ dt \cdot dW_i = 0, i = 1, \ldots, n, \\ dW_i \cdot dW_j = \rho_{i,j}dt. \end{cases} \tag{1.22}$$

1.3. El Modelo de Black-Scholes

Se considera un mercado financiero en el que existen dos activos:

Un activo libre de riesgo B, que sigue un proceso determinista dado por la dinámica

$$dB(t) = rB(t)dt\,, \tag{1.23}$$

y una acción S, que sigue un proceso estocástico dado por la dinámica

$$dS(t) = \mu S(t)dt + \sigma S(t)dW(t)\,, \tag{1.24}$$

donde W es un proceso de Wiener en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$, y recordemos que P es la medida física.

Fischer Black y Myron Scholes, en su famoso artículo [7], asumen además las siguientes hipótesis:

1. No existen oportunidades de arbitraje.
2. Se permiten las posiciones largas y cortas de cualquiera de los activos.
3. Está permitido comprar o vender una cantidad ilimitada de cualquier activo en el mercado.
4. El precio de venta de un activo es igual a su precio de compra.
5. No existen costes de transacción.
6. La acción no paga dividendos.
7. La negociación de los activos se lleva a cabo de manera continua.

Aunque en la práctica pueden existir oportunidades de arbitraje, si lo hacen, estas existen por muy poco tiempo.

También se podría considerar un modelo donde la acción pagara dividendos o incluyera los costes de transacción, pero en general, las hipótesis que se toman, aunque simplifican mucho la realidad, permiten obtener a cambio una fórmula explícita para el precio de ciertos derivados financieros.

La forma en que se va a valorar la opción europea se basa en la construcción de lo que se llama un portfolio de réplica.

Definición 1.24. Dado el mercado formado por el bono y la acción, se denomina *portfolio de réplica* de la Call Europea a la estrategia $[\alpha_0(t), \alpha_1(t)]$ y el portfolio de inversiones $\Pi(t)$ dado por

$$\Pi(t) = \alpha_0(t)B(t) + \alpha_1(t)S(t)\,, \qquad t \in [0, T]\,,$$

que verifica que $\Pi(T) = \text{máx}(S - K, 0)$.

Se dice que el portfolio es *autofinanciado* si

$$d\Pi(t) = \alpha_0(t)dB(t) + \alpha_1(t)dS(t)\,.$$

Nota. La propiedad de autofinanciamiento implica que el portfolio puede rebalancearse sin la necesidad de aportes adicionales de capital.

En el caso de que existiera un portfolio de réplica autofinanciado, el precio de la Call Europea y el del portfolio deberían coincidir en todo momento ya que, en caso contrario, surgiría una oportunidad de arbitraje comprando/vendiendo la Call y vendiendo/comprando el portfolio de forma similar a como se vio en la Proposición 1.2.

La demostración de la existencia de este tipo de portfolios es bastante extensa y técnica (ver [6]), necesitando de resultados de Teoría de Probabilidad y Análisis Funcional que están más allá del objetivo de este trabajo. No obstante, se presenta un breve resumen de las etapas intermedias de la demostración.

Definición 1.25. Sea $(\Omega, \mathcal{F})$ un espacio medible y sean P y Q dos medidas sobre $(\Omega, \mathcal{F})$.

Se dice que P y Q son *medidas equivalentes* y se denota $P \sim Q$, si se cumple que

$$P(A) = 0 \iff Q(A) = 0\,, \tag{1.25}$$

es decir, si las dos medidas tienen exactamente los mismos conjuntos de medida nula.

De forma general (ver capítulo 10 de [6]), se considera un mercado sobre el espacio de probabilidad $(\Omega, \mathcal{F}, P)$, formado por los activos $S_0, S_1, \ldots, S_N$, que siguen dinámicas dadas por

$$dS_i(t) = \mu_i(t)S_i(t)dt + \sigma_i(t)S_i(t)dW\,,$$

y sea $\Pi(t; X)$ un portfolio replicador para el derivado X que se ejecuta en el vencimiento y cuyo valor dependerá de los valores $S_0, S_1, \ldots, S_N$.

Definición 1.26. Se dice que Q es una *medida libre de riesgo* o *medida de martingala*, si $Q \sim P$ y los procesos

$$\frac{S_0(t)}{S_0(t)}, \frac{S_1(t)}{S_0(t)}, \ldots \frac{S_N(t)}{S_0(t)}\,,$$

son martingalas sobre Q.

El activo S_0 se denomina habitualmente el activo numerario del mercado.

Teorema 1.27 (Primer Teorema Fundamental)**.** *El modelo del mercado esta libre de arbitraje si, y solamente si, existe una medida libre de riesgo.*

Proposición 1.28. *Si el activo numerario S_0 verifica que $\sigma_0(t) \equiv 0$, es decir,*

$$S_0(t) = e^{\int_0^t r(s)ds},$$

entonces una medida Q equivalente a P es una medida libre de riesgo si, y solamente si los activos $S_1 \ldots, S_N$ tienen Q-dinámicas en la medida Q de la forma

$$dS_i(t) = S_i(t)r(t)dt + S_i(t)\sigma_i(t)dW^Q, \qquad i = 1 \ldots, N.$$

donde W^Q es un proceso de Wiener multidimensional en la medida Q.

Teorema 1.29 (Segundo Teorema Fundamental)**.** *Suponiendo que el mercado esta libre de arbitraje, el mercado es completo (i.e todo derivado se puede replicar mediante un portfolio replicador autofinanciado) si, y solo si la medida libre de riesgo es única.*

Proposición 1.30. *Si $\Pi(t; X)$ es el precio del portfolio replicador de un derivado financiero X como el indicado, se verifica que*

1. *Para evitar oportunidades de arbitraje, el precio de X tiene que venir dado por*

$$\Pi(t; X) = S_0(t)\mathbb{E}^Q\left[\frac{X}{S_0(t)}\Big|\mathcal{F}_t\right],$$

   donde Q es la medida libre de riesgo para $[S_0, S_1, \ldots, S_N]$ y S_0 es el activo numerario.

2. *En particular, si el activo numerario es de la forma*

$$dS_0(t) = r(t)S_0(t)dt,$$

   la fórmula de valoración anterior es

$$\Pi(t; X) = \mathbb{E}^Q\left[e^{-\int_t^T r(s)ds}X\Big|\mathcal{F}_t\right].$$

Se dice que el mercado *sigue el modelo de Black-Scholes* si las dinamicas de los activos vienen dadas por

$$\begin{cases} dB(t) = rB(t)dt \\ dS(t) = \mu S(t)dt + \sigma S(t)dW(t) \end{cases}, \tag{1.26}$$

donde r, μ y σ son constantes.

Por tanto, para el modelo de Black-Scholes, si existiera la medida libre de riesgo Q, se tendría el siguiente resultado.

Teorema 1.31 (Valoración libre de Riesgo)**.** *El precio $C(t)$ para el modelo de Black-Scholes de una Call Europea sobre el activo subyacente S, de strike K y vencimiento T,viene dado en cada $t \in [0,T]$ por:*

$$C(t) = e^{r(T-t)}\mathbb{E}^Q[\text{máx}(S(T)-K,0)|\mathcal{F}^{W^Q}]\,. \tag{1.27}$$

y la dinámica de las acciones del Modelo de Black-Scholes viene dada por:

$$dS(t) = rS(t)dt + \sigma S(t)dW^Q(t)\,. \tag{1.28}$$

Entonces, para concluir la sección, indicar que a partir del Teorema de Girsanov (ver [6]), se demuestra que el cambio de variable dado por $dQ = L_T dP$, donde

$$L_T = \exp\Big(\int_0^T \varphi(s)dW(s) - \frac{1}{2}\int_0^T |\varphi(s)|^2 ds\Big)\,.$$

con $\varphi = \frac{\mu - r}{\sigma}$, lleva a la media Q buscada.

1.3.1. La Fórmula de Black-Scholes

Lema 1.32. *Sea $X \sim \mathcal{N}(\mu,\sigma)$ y sea $\Phi(x)$ la función de distribución de X. Para todo $x \in \mathbb{R}$ se verifica que:*

$$\Phi(-x) = 1 - \Phi(x)\,. \tag{1.29}$$

Demostración. Por definición se tiene que

$$\Phi(-x) = \int_{-\infty}^{-x} f_X(t)dt\,,$$

donde f_X es la función de densidad de la distribución normal definida en (1.5).

En virtud del Teorema de Cambio de Variable (ver [15]), tomando $t = -u$, se llega a que:

$$\Phi(-x) = -\int_{\infty}^{x} f_X(u)du = \int_x^{\infty} f_X(u)du\,. \tag{1.30}$$

Puesto que f_X es función de densidad de una variable aleatoria, se cumple que

$$\int_{-\infty}^{\infty} f_X(u)du = 1\,.$$

Entonces, como para cada $x \in \mathbb{R}$, se tiene que

$$\int_{-\infty}^{\infty} f_X(u)du = \int_{-\infty}^{x} f(u)du + \int_{x}^{\infty} f(u)du = 1\,,$$

despejando y aplicando la definición de función de distribución, se llega a que:

$$\int_{x}^{\infty} f(u)du = 1 - \Phi(x)\,.$$

Juntando la expresión anterior con la de (1.30), se concluye que

$$\Phi(-x) = 1 - \Phi(x)\,,$$

como se quería demostrar. □

Lema 1.33. *Sea $X \sim \mathcal{N}(\mu, \sigma)$ y sean $a, k \in \mathbb{R}$ constantes. Entonces se verifica que:*

$$\mathbb{E}[\exp(aX)|X \geq k] = \exp\Big(a\mu + \frac{a^2\sigma^2}{2}\Big)N(d)\,, \tag{1.31}$$

donde $d = \frac{-k+\mu+a\sigma^2}{\sigma}$ y N es la función de distribución de una variable aleatoria normal estándar.

Demostración. Por la definición de esperanza condicionada y en virtud de las propiedades que verifica la esperanza de una variable aleatoria, se tiene que:

$$\mathbb{E}[\exp(aX)|X \geq k] = \frac{1}{\sigma\sqrt{2\pi}} \int_{k}^{\infty} \exp(ax) \exp\Big(-\frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big)dx\,. \tag{1.32}$$

donde aplicando las propiedades de las exponenciales, se obtiene que

$$\mathbb{E}[\exp(aX)|X \geq k] = \frac{1}{\sigma\sqrt{2\pi}} \int_{k}^{\infty} \exp\Big(ax - \frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\Big)dx\,.$$

Desarrollando el argumento de la función exponencial y completando cuadrados se llega a

$$\begin{aligned} ax - \frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2 &= \frac{2\sigma^2 ax - x^2 - \mu^2 + 2x\mu}{2\sigma^2} = \frac{-x^2 + 2x(\mu + a\sigma^2) - \mu^2}{2\sigma^2} \\ &= -\frac{x^2 - 2x(\mu + a\sigma^2) + (\mu + a\sigma^2)^2 - (\mu + a\sigma^2)^2 + \mu^2}{2\sigma^2} \\ &= \frac{(x - (\mu + a\sigma^2))^2 - a^2\sigma^4 - 2a\sigma^2\mu}{2\sigma^2} \\ &= a\mu + \frac{a^2\sigma^2}{2} - \frac{1}{2}\Big(\frac{x - (\mu + a\sigma^2)}{\sigma}\Big)^2\,. \end{aligned}$$

Por lo tanto, se tiene que

$$\begin{aligned}\mathbb{E}[\exp(aX)|X \geq k] &= \frac{1}{\sigma\sqrt{2\pi}} \int_k^\infty \exp\Big(a\mu + \frac{a^2\sigma^2}{2} - \frac{1}{2}\Big(\frac{x-(\mu+a\sigma^2)}{\sigma}\Big)^2\Big)dx \\ &= \exp\Big(a\mu + \frac{a^2\sigma^2}{2}\Big)\frac{1}{\sigma\sqrt{2\pi}} \int_k^\infty \exp\Big(\frac{1}{2}\Big(\frac{x-(\mu+a\sigma^2)}{\sigma}\Big)^2\Big)dx\end{aligned}$$

Considerando el cambio de variable $u = \frac{x-(\mu+a\sigma^2)}{\sigma}$ en la expresión anterior, el límite superior de integración se mantiene igual mientras que el inferior, pasa a ser $L = \frac{k-(\mu+a\sigma^2)}{\sigma}$. Además, se tiene que $du = \frac{1}{\sigma}dx$, con lo cual, tras aplicar el Teorema del Cambio de Variable, se tiene que:

$$\mathbb{E}[\exp(aX)|X \geq k] = \exp\Big(a\mu + \frac{a^2\sigma^2}{2}\Big)\frac{1}{\sqrt{2\pi}} \int_L^\infty \exp\Big(-\frac{u^2}{2}\Big)du\,. \tag{1.33}$$

Por definición,

$$N(L) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^L \exp\Big(-\frac{u^2}{2}\Big)du\,,$$

entonces, por las propiedades de la función de distribución y aplicando el Lema 1.32, se tiene que

$$\frac{1}{\sqrt{2\pi}} \int_L^\infty \exp\Big(-\frac{u^2}{2}\Big)du = 1 - N(L) = N(-L)\,.$$

Con lo cual, llamando $d = -L = \frac{-k+\mu+a\sigma^2}{\sigma}$, se puede reescribir (1.33) como

$$\mathbb{E}[\exp(aX)|X \geq k] = \exp\Big(a\mu + \frac{a^2\sigma^2}{2}\Big)N(d)\,,$$

que es lo que se quería demostrar. □

Teorema 1.34 (Fórmula de Black-Scholes para la Call Europea.)**.** *El precio C de una Call Europea sobre el activo subyacente S, de strike K y vencimiento T,viene dado en cada $t \in [0,T]$ por:*

$$C(t) = S(t)N(d_1) - Ke^{-r(T-t)}N(d_2)\,, \tag{1.34}$$

donde:

$$d_1 = \frac{\ln\Big(\frac{S(t)}{K}\Big) + \Big(r + \frac{\sigma^2}{2}\Big)(T-t)}{\sigma\sqrt{T-t}}\,, d_2 = \frac{\ln\Big(\frac{S(t)}{K}\Big) + \Big(r - \frac{\sigma^2}{2}\Big)(T-t)}{\sigma\sqrt{T-t}}\,,$$

- N: *es la función de distribución de una variable aleatoria* $\mathcal{N}(0,1)$,
- K: *es el strike o precio de ejercicio,*
- T: *es la fecha de vencimiento,*
- t: *es el instante de la valoración de la opción,*
- $S(t)$: *es el precio de la opción en el instante* t,
- r: *es el tipo de interés del activo libre de riesgo y*
- σ: *es la volatilidad que presenta la dinámica de precios del activo con riesgo.*

Demostración. Por el Teorema 1.31, se tiene que el precio de una Call Europea en un instante t previo al vencimiento es

$$C(t) = e^{-r(T-t)}\mathbb{E}^Q[\text{máx}(S(T)-K,0)|\mathcal{F}_t^{W^Q}]\,, \tag{1.35}$$

luego, la Call se ejecuta si $S(T) \geq K$ y vale 0 en caso contrario.

Puesto que la acción sigue la dinámica dada por el proceso de Itô

$$dS(t) = rS(t)dt + \sigma S(t)dW(t)\,,$$

aplicando el Lema de Itô para la función $f(t,x) = \ln(x)$, como

$$\frac{\partial f}{\partial t}(t,x) = 0\,, \frac{\partial f}{\partial x}(t,x) = \frac{1}{x} \text{ y } \frac{\partial^2 f}{\partial x^2}(t,x) = -\frac{1}{x^2}\,,$$

sustituyendo en (1.12) se llega a que:

$$d\ln(S(t)) = \Big(r - \frac{\sigma^2}{2}\Big)dt + \sigma dW(t)\,. \tag{1.36}$$

por lo que

$$\ln\Big(\frac{S(T)}{S(t)}\Big) = \Big(r - \frac{\sigma^2}{2}\Big)(T-t) - \sigma(W(T)-W(t))\,.$$

Para que la Call se ejecute se tiene que cumplir que

$$S(T) = S(t)\exp\Big(\Big(r - \frac{\sigma^2}{2}\Big)(T-t) - \sigma(W(T)-W(t))\Big) \geq K\,, \tag{1.37}$$

lo que equivale a

$$(W(T)-W(t)) \geq \frac{1}{\sigma}\Big(\ln\Big(\frac{K}{S(t)}\Big) - \Big(r - \frac{\sigma^2}{2}\Big)(T-t)\Big)\,.$$

Tomando $k = \frac{1}{\sigma}\Big(\ln\Big(\frac{K}{S(t)}\Big) - \Big(r - \frac{\sigma^2}{2}\Big)(T-t)\Big)$, como la opción vale 0 si no se ejecuta, y de hacerlo su precio es

$$C(t) = e^{-r(T-t)}\mathbb{E}\Big[S(t)\exp\Big(\Big(r-\frac{\sigma^2}{2}\Big)(T-t) - \sigma(W(T)-W(t))\Big) - K\Big|W(T)-W(t) \geq k\Big]\,,$$

Aplicando las propiedades de la esperanza se tiene que

$$\begin{aligned} C(t) = S(t)\exp\Big(-\frac{\sigma^2(T-t)}{2}\Big)\mathbb{E}\Big[\exp\Big(\sigma(W(T)-W(t))\Big)\Big|W(T)-W(t) \geq k\Big] \\ - Ke^{-r(T-t)}\mathbb{E}[1|W(T)-W(t) \geq k]\,, \end{aligned}$$

y aplicando en ambos términos el Lema 1.33 ($1 = \exp(0)$), tras las cuentas oportunas se llega al resultado deseado.

□

1.3.2. La EDP de Black-Scholes

En la sección anterior, se ha obtenido una fórmula para el precio de una Call Europea, proporcionando una herramienta directa para determinar el precio de este tipo específico de opción. A continuación, se expondrá la derivación de la EDP de Black-Scholes y su aplicación para la valoración de derivados.

Sea $V(S,t)$ el precio en el instante t de un derivado financiero con strike K y vencimiento T sobre un activo subyacente S que sigue la dinámica de (1.24).

Sea Π el valor de un portfolio compuesto por una posición larga del derivado y por una cantidad Δ de la acción.

Entonces, el valor del portfolio en el instante t es:

$$\Pi(S,t) = V(S,t) - \Delta S\,. \tag{1.38}$$

Si se considera la variación del valor del portfolio en un instante, se tiene que:

$$d\Pi = dV - \Delta dS\,. \tag{1.39}$$

Por las hipótesis del modelo de Black-Scholes que asumimos, se tiene que el proceso S sigue la dinámica

$$dS = \mu S dt + \sigma S dW\,, \tag{1.40}$$

luego, en virtud del Lema de Itô, se tiene que:

$$dV = \Big(\frac{\partial V}{\partial t} + \mu S\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}\Big)dt + \sigma S\frac{\partial V}{\partial S}dW\,. \tag{1.41}$$

Sustituyendo la expresión anterior en (1.39), se llega a que

$$d\Pi = \Big(\frac{\partial V}{\partial t} + \mu S\frac{\partial V}{\partial S} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2} - \mu S\Delta\Big)dt + \sigma S\Big(\frac{\partial V}{\partial S} - \Delta\Big)dW\,. \tag{1.42}$$

Se advierte que el primer término de la expresión anterior es determinista, mientras que el segundo término tiene un carácter aleatorio. Entonces, tomando

$$\Delta = \frac{\partial V}{\partial S}\,, \tag{1.43}$$

se elimina la componente aleatoria de la dinámica del portfolio Π

En consecuencia, dado que la evolución del portfolio es completamente determinista, el portfolio debe de evolucionar de la misma manera que el activo libre de riesgo del mercado. Por lo que, en virtud de la Proposición 1.1, debe cumplirse que:

$$d\Pi = r\Pi dt\,, \tag{1.44}$$

Entonces, puesto que necesariamente se tiene que verificar (1.44) ya que de lo contrario, habría una oportunidad de arbitraje, imponiendo la definición del portfolio de (1.38) y la expresión de (1.42) teniendo en cuenta el Δ que se ha tomado, se llega a que

$$\Big(\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2}\Big)dt = r\Big(V - S\frac{\partial V}{\partial S}\Big)dt\,, \tag{1.45}$$

donde reordenando los términos, se obtiene la conocida Ecuación de Black-Scholes:

$$\boxed{\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2\frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0} \tag{1.46}$$

Nota. A la práctica de escoger la cantidad Δ como en (1.43) para eliminar el riesgo del portfolio se la conoce como *delta hedging*.

Para que la ecuación quede totalmente determinada, falta imponer las condiciones frontera en la ecuación.

Por ejemplo, si se supone que $V(S,t) = C(S,t)$ es el valor de una Call Europea, con strike K y vencimiento T, el valor en el vencimiento debe ser

$$C(S,T) = \text{máx}(S-K,0)\,. \tag{1.47}$$

Si $S = 0$, debido a la dinámica de la acción (1.24), S valdrá 0 hasta el vencimiento, luego, se tiene que verificar que

$$C(0,t) = 0, \qquad t \geq 0\,. \tag{1.48}$$

Si se considera $P(S,t)$ como es el valor de una Call Europea con strike K y vencimiento T, cuando S tiende a infinito, será más difícil que se ejerza y, en el límite se tiene con probabilidad probabilidad 1 que

$$\lim_{S\to\infty} P(S,t) = 0\,, \qquad t \geq 0\,,$$

por lo que, en virtud de la Proposición 1.2

$$\lim_{S\to\infty} \frac{C(S,t)}{S} = \frac{P(S,t) - Ke^{-r(T-t)} + S}{S} = 1\,, \qquad t \geq 0\,. \tag{1.49}$$

Es posible obtener la fórmula de (1.34) mediante la resolución de la EDP (1.46) imponiendo las condiciones frontera (1.47), (1.48) y (1.49). Este enfoque se puede consultar en [37].

Para derivados más exóticos (opciones americanas, barrera, etc.) u otras dinámicas en general, no se puede obtener una fórmula explícita para determinar su precio. Sin embargo, la EDP de Black-Scholes permite utilizar métodos numéricos de resolución de ecuaciones diferenciales en derivadas parciales, como las diferencias finitas y los elementos finitos.

1.3.3. Deficiencias del Modelo de Black-Scholes

El modelo de Black-Scholes supuso un hito en la valoración de derivados financieros, pero este modelo presenta una serie de carencias debido a las hipótesis que asume para simplificar la realidad de los mercados financieros.

La principal de estas limitaciones es la suposición de que tanto la volatilidad como la tasa libre de riesgo son constantes en el tiempo, cosa que, como se muestra en [16], no se ajusta con las observaciones empíricas que se realizan en el mercado.

El modelo de Black-Scholes tampoco tiene en cuenta los costes de transacción que aparecen en la práctica. De hecho, la estrategia de cobertura *delta-hedging*

descrita anteriormente, se basa en mantener continuamente una cantidad $\frac{\partial V}{\partial S}$ del activo subyacente. Debido a que la estrategia se realiza de manera continua, en [36] se prueba que si existen los costes de transacción , la ejecución de una estrategia de cobertura mediante el portfolio replicador que se construye mediante *delta-hedging* conlleva la pérdida de todo el capital de la inversión, siendo la única estrategia viables comprar el propio subyacente.

La suposición que hace el modelo de Black-Scholes sobre la liquidez del mercado, es decir, sobre la posibilidad de poder comprar o vender cualquier cantidad de cualquier activo en el mercado tampoco resulta realista en la práctica.

Sin embargo, pese a todos estos inconvenientes, el modelo de Black-Scholes ha sido el principal marco de referencia para la valoración de derivados financieros durante muchos años y, en la actualidad, se sigue usando en la práctica en muchas ocasiones debido a que se dispone de una fórmula explícita fácil de evaluar.

1.4. El Modelo de Heston

Para superar algunas limitaciones del modelo de Black-Scholes, como la asunción de volatilidad constante, surgieron otros modelos más sofisticados, que por ejemplo asumen que la volatilidad es otro proceso estocástico.

De esta serie de modelos de volatilidad estocástica en tiempo continuo, destaca el introducido en 1993 por Steven L. Heston en [19], ya que proporciona una fórmula semi-explícita para el precio de una Call Europea.

En su modelo, Heston asume que la dinámica del activo con riesgo viene dada por

$$\begin{cases} dS(t) = \mu S(t)dt + \sqrt{v(t)}S(t)dW_1(t), & (1.50) \\ dv(t) = \kappa[\theta - v(t)]dt + \sigma\sqrt{v(t)}dW_2(t), & (1.51) \end{cases}$$

donde W_1 y W_2 son dos procesos de Wiener cuya correlación viene dada por $\rho = \mathrm{Cov}(W_1, W_2)$ y la dinámica $v(t)$ de la varianza del proceso de la acción que se define en (1.51) viene dada por un proceso de Ornstein–Uhlenbeck de reversión a la media. El parámetro θ representa el valor en torno a el cual el proceso oscila, κ es la tasa de reversión y σ es la "volatilidad de la volatilidad".

Mediante el Lema de Itô multidimensional y empleando argumentos similares s a los utilizados para deducir (1.46) , se demuestra en [16] que el valor de un derivado financiero debe de satisfacer

$$\frac{1}{2}vS^2\frac{\partial^2 V}{\partial S^2} + \rho\sigma vS\frac{\partial^2 V}{\partial S \partial v} + \frac{1}{2}\sigma^2 v\frac{\partial^2 V}{\partial v^2} + r_0 S\frac{\partial V}{\partial S} + (\kappa(\theta - v) - \lambda(S, v, t))\frac{\partial V}{\partial v} - rV + \frac{\partial V}{\partial t} = 0\,,$$

donde $\lambda(S,v,t)$ representa el precio de la prima de riesgo.

Como se indica en [19], considerando argumentos del modelo de consumo de Cox, Ingersoll y Ross (ver [11]) la prima de riesgo es $\lambda(S,v,t)=\lambda v$. Por tanto, la EDP que satisface el precio $V(S,v,t)$ de una Call Europea es

$$\frac{1}{2}vS^2\frac{\partial^2 V}{\partial S^2}+\rho\sigma vS\frac{\partial^2 V}{\partial S\partial v}+\frac{1}{2}\sigma^2 v\frac{\partial^2 V}{\partial v^2}+rS\frac{\partial V}{\partial S}+(\kappa(\theta-v)-\lambda v)\frac{\partial V}{\partial v}-r_0V+\frac{\partial V}{\partial t}=0\,, \tag{1.52}$$

sujeta las condiciones frontera

$$\begin{aligned} V(S,v,T) &= \text{máx}(S-K,0),\\ V(0,v,t) &= 0,\\ r_S\frac{\partial V}{\partial S}(S,0,t)+\kappa\theta\frac{\partial V}{\partial v}(S,0,t)-rV(S,0,t)+\frac{\partial V}{\partial t}(S,0,t) &= 0,\\ \frac{\partial V}{\partial S}(\infty,v,t) &= 1,\\ V(S,\infty,t) &= S. \end{aligned}$$

A partir del Teorema 1.31, se tiene que el precio de una Call Europea en la medida libre de riesgo Q tiene que verificar

$$\begin{aligned} V(S,v,t) &= e^{-r(T-t)}\mathbb{E}^Q[\text{máx}(S(T)-K,0)|\mathcal{F}_t] =\\ &= e^{-r(T-t)}\mathbb{E}^Q[(S(T)-K)\mathbb{1}_{\{S(T)\geq K\}}|\mathcal{F}_t] =\\ &= \mathbb{E}^Q[e^{-r(T-t)}S(T)\mathbb{1}_{\{S(T)\geq K\}}|\mathcal{F}_t]-Ke^{-r(T-t)}\mathbb{E}^Q[\mathbb{1}_{\{S(T)\geq K\}}|\mathcal{F}_t] =\\ &= S(t)\mathbb{E}^Q\left[\frac{e^{-r(T-t)}S(T)}{S(t)}\mathbb{1}_{\{S(T)\geq K\}}\middle|\mathcal{F}_t\right]-Ke^{-r(T-t)}\mathbb{E}^Q[\mathbb{1}_{\{S(T)\geq K\}}|\mathcal{F}_t]\,, \end{aligned}$$

por lo que se puede asumir que existe una solución similar a la de Black-Scholes

$$V(S,v,t)=SP_1(S,v,t)-Ke^{-r(T-t)}P_2(S,v,t)\,, \tag{1.53}$$

donde

$$\begin{cases} P_1(S,v,t)=\mathbb{E}^Q[\frac{e^{-r(T-t)}S(T)}{S(t)}\mathbb{1}_{\{S(T)\geq K\}}|\mathcal{F}_t]\,,\\ P_2(S,v,t)=\mathbb{E}^Q[\mathbb{1}_{\{S(T)\geq K\}}|\mathcal{F}_t]\,. \end{cases}$$

Se tiene que $e^{-r(T-t)}P_2(S,v,t)$ es el precio de un derivado cuyo valor en el vencimiento viene determinado por $\mathbb{1}_{\{S(T)\geq K\}}$. Entonces, $P_2(S,v,t)$ debe verificar la ecuación (1.52).

Realizando el cambio de variable $x = \ln(S)$ se tiene que $P_2(x, v, t)$ verifica

$$\frac{1}{2}v\frac{\partial^2 P_2}{\partial x^2} + \rho\sigma v\frac{\partial^2 P_2}{\partial x \partial v} + \frac{1}{2}\sigma^2 v\frac{\partial^2 P_2}{\partial v^2} + \left(r - \frac{v}{2}\right)\frac{\partial P_2}{\partial x} + (\kappa(\theta - v) - \lambda v)\frac{\partial P_2}{\partial v} + \frac{\partial P_2}{\partial t} = 0\,, \tag{1.54}$$

con la condición frontera $P_2(x, v, T) = \mathbb{1}_{\{x \geq \ln(K)\}}$.

Sustituyendo (1.53) en (1.52) y teniendo en cuenta que P_2 verifica (1.54), se obtiene que $P_1(x, v, t)$ verifica

$$\frac{1}{2}v\frac{\partial^2 P_1}{\partial x^2} + \rho\sigma v\frac{\partial^2 P_1}{\partial x \partial v} + \frac{1}{2}\sigma^2 v\frac{\partial^2 P_1}{\partial v^2} + (r + \frac{v}{2})\frac{\partial P_1}{\partial x} + (\kappa(\theta - v) - \lambda v + \rho\sigma v)\frac{\partial P_1}{\partial v} + \frac{\partial P_1}{\partial t} = 0\,, \tag{1.55}$$

satisfaciendo también la condición frontera $P_1(x, v, T) = \mathbb{1}_{\{x \geq \ln(K)\}}$.

Por simplicidad, para $j = 1, 2$ se tiene que $P_j(S, v, T)$ verifica

$$\frac{1}{2}v\frac{\partial^2 P_j}{\partial x^2} + \rho\sigma v\frac{\partial^2 P_j}{\partial x \partial v} + \frac{1}{2}\sigma^2 v\frac{\partial^2 P_j}{\partial v^2} + (r + u_j v)\frac{\partial P_j}{\partial x} + (a_j - b_j v)\frac{\partial P_j}{\partial v} + \frac{\partial P_j}{\partial t} = 0\,, \tag{1.56}$$

sujeto a $P_j(x, v, T) = \mathbb{1}_{\{x \geq \ln(K)\}}$, donde

$$u_1 = \frac{1}{2},\ u_2 = -\frac{1}{2},\ a_1 = a_2 = \kappa\theta,\ b_1 = (\kappa + \lambda) - \rho\sigma,\ b_2 = (\kappa + \lambda)\,. \tag{1.57}$$

Para cada $j = 1, 2$, tras los cambios de variables indicados, se consideran los procesos estocásticos x_j e v_j dados por

$$\begin{cases} dx_j(t) = (r + u_j v(t))dt + \sqrt{v(t)}dW_1(t), & (1.58) \\ dv_j(t) = (a_j - b_j v(t))dt + \sigma\sqrt{v(t)}dW_2(t), & (1.59) \end{cases}$$

y se tiene que

$$P_j(x, v, t; \ln(K)) = \mathbb{P}\Big(x_j(T) \geq \ln(K) | x_j(t) = x, v_j(t) = v\Big)\,. \tag{1.60}$$

Esto se ve definiendo, para cada $j = 1, 2$,

$$g_j(x, v, t; \ln(K)) = \mathbb{E}[\mathbb{1}_{\{x_j(T) \geq \ln(K)\}} | x_j(t) = x, v_j(t) = v]\,,$$

y viendo, mediante el Teorema de la Esperanza Total (ver [5]), que g_j es una martingala, ya que para $s \leq t$,

$$\begin{aligned} \mathbb{E}[g_j(t)|\mathcal{F}_s] &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{x_j(T) \geq \ln(K)\}} | x_j(t) = x, v_j(t) = v]|\mathcal{F}_s] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{1}_{\{x_j(T) \geq \ln(K)\}}|\mathcal{F}_t]|\mathcal{F}_s] = \mathbb{E}[g_j(t)|\mathcal{F}_s] \\ &= \mathbb{E}[\mathbb{1}_{\{x_j(T) \geq \ln(K)\}}|\mathcal{F}_s] \\ &= \mathbb{E}[\mathbb{1}_{\{x_j(T) \geq \ln(K)\}} | x_j(s) = x, v_j(s) = v] = g_j(s)\,. \end{aligned}$$

Aplicando el Lema de Itô multidimensional y la Proposición 1.20 se tiene, tras las oportunas cuentas, que g_j es solución de la EDP

$$\begin{aligned}&\frac{\partial g_j}{\partial t} + (r + u_j v_j(t))\frac{\partial g_j}{\partial x} + (a_j - b_j v_j(t))\frac{\partial g_j}{\partial v} + \frac{1}{2}v_j(t)\frac{\partial^2 g_j}{\partial x^2} + \\ &+ \frac{1}{2}\sigma^2 v_j(t)\frac{\partial^2 g_j}{\partial v^2} + \rho\sigma v_j(t)\frac{\partial^2 g_j}{\partial x \partial v} = 0\,,\end{aligned} \tag{1.61}$$

y que, por definición de g_j se tiene que

$$g_j(x, v, T; \ln(K)) = \mathbb{1}_{\{x \geq \ln(K)\}}\,.$$

Entonces, para cada $j = 1, 2$, g_j es solución de (1.56), y se tiene que

$$P_j(x, v, t; \ln(K)) = g_j(x, v, t; \ln(K))\,.$$

probando la afirmación realizada en (1.60).

Para calcular las probabilidades de (1.60), para cada $j = 1, 2$ se acude a las funciones características de x_j

$$f_j(\phi; x, v, t) = \mathbb{E}[e^{i\phi x_j(T)} | x_j(t) = x, v_j(t) = v]\,,$$

Se asume (ver [19]), que la función característica es de la forma

$$f_j(\phi; x, v, t) = \exp(C_j(T - t; \phi) + D_j(T - t; \phi)v + i\phi x)\,. \tag{1.62}$$

donde $f_j(\phi; x, v, T) = \exp(i\phi x)$ y donde $C_j(\phi; T-t)$ y $D_j(\phi; T-t)$ son dos funciones a determinar.

Aplicando el lema de Itô y con argumentos similares a los utilizados, f_j satisface (1.61) sujeta a la condición $f_j(\phi; x, v, T) = \exp(i\phi x)$.

Sustituyendo la solución de (1.62) en la ecuación de (1.61) y reordenando los términos, se obtiene que las funciones C_j y D_j satisfacen las siguientes ecuaciones diferenciales ordinarias

$$\begin{cases} -\frac{1}{2}\sigma^2\phi^2 + \rho\sigma\phi i D_j + \frac{1}{2}D_j^2 + u_j\phi i - b_j D_j + \frac{\partial D_j}{\partial t} = 0, & (1.63) \\ r\phi i + a_j D_j + \frac{\partial C_j}{\partial t} = 0, & (1.64) \end{cases}$$

sujetas a las condiciones iniciales $C_j(0; \phi) = D_j(0; \phi) = 0$.

La EDO que aparece en (1.63) es una Ecuación de Riccati (ver [28]) cuya solución es

$$D_j(\tau;\phi) = \frac{b_j - \rho\sigma\phi i + d}{\sigma^2}\Big(\frac{1-e^{d\tau}}{1-ge^{d\tau}}\Big)\,, \tag{1.65}$$

donde

$$\tau = T-t,\ d = \sqrt{(\rho\sigma\phi i - b_j)^2 - \sigma^2(2u_j\phi i - \phi^2)}\ \text{y}\ g = \frac{b_j - \rho\sigma\phi i + d}{b_j - \rho\sigma\phi i - d}\,.$$

Conocido D_j e integrando (1.64) para despejar C_j, imponiendo la condición $C_j(0;\phi) = 0$ se obtiene que

$$C_j(\tau;\phi) = r\phi i\tau + \frac{a_j}{\sigma^2}\Big[(b_j - \rho\sigma\phi i + d)\tau - 2\ln\Big(\frac{1-ge^{d\tau}}{1-g}\Big)\Big]\,, \tag{1.66}$$

con lo que queda perfectamente determinada la función $f_j(\phi;x,v,t)$.

Para calcular (1.60), simplemente basta con invertir las funciones características aplicando el Teorema 1.10 para obtener la probabilidad deseada

$$P_j(x,v,t;\ln(K)) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \mathrm{Re}\Big(\frac{e^{-i\phi\ln(K)}f_j(\phi;x,v,t)}{i\phi}\Big)d\phi\,, \tag{1.67}$$

para cada $j = 1, 2$.

Con lo cual, el precio que da el modelo de Heston para la Call es

$$\boxed{V(S,v,t;K,T,r,\kappa,\theta,\sigma,\lambda) = SP_1(\ln(S),v,t) - Ke^{-r(T-t)}P_2(\ln(S),v,t)\,.} \tag{1.68}$$

1.4.1. Real Time Trade

Como se ha visto, el modelo de Black-Scholes proporciona una fórmula explícita para el precio de la Call Europea. No obstante, los modelos que ofrecen una fórmula cerrada para el precio de un derivado son escasos. Ya se ha visto que un modelo algo más sofisticado (Heston) solo ofrece una fórmula semi-cerrada y, en general, cuando se valoran derivados más complejos (opciones barrera, asiáticas, etc.), se hace imperativo el uso de métodos numéricos para aproximar los precios.

Existen diferentes técnicas para estimar el precio de un derivado. Si el modelo trabaja con la EDP, existen diversos métodos para su solución numérica. Por ejemplo, se puede emplear el método de elementos finitos como se describe en [1], o se pueden utilizar métodos espectrales, cuya aplicación en la valoración de derivados financieros se detalla en [24].

También es posible el uso de técnicas estadísticas para la valoración de un derivado. El Teorema 1.31 ofrece una fórmula (generalizable para cualquier derivado) que permite aplicar simulaciones de Montecarlo para aproximar el valor de derivados. En [18] se presenta una descripción exhaustiva de los métodos de Montecarlo en el ámbito de las finanzas, destacando el Algoritmo de Longstaff-Schwartz, [25], para las valoración de opciones americanas.

En la práctica, es importante tener la capacidad de obtener el precio de un derivado de manera (casi) inmediata tras un cambio en las condiciones del mercado (un cambio en la cotización, negociación de un nuevo vencimiento, etc.). A este problema se le conoce como *Real Time Trade* o valoración en tiempo real. En este contexto, es fundamental la rapidez de los métodos numéricos que se empleen para la valoración, ya que para ejecutar órdenes de compra-venta, es necesario determinar el precio del derivado antes de que la cotización de su subyacente cambie de nuevo, cosa que sucede de manera continua en los mercados financieros,

En este contexto, las redes neuronales representan una opción interesante para abordar el problema de la valoración en tiempo real ya que, como se verá en el siguiente capítulo, el coste computacional de evaluar una red previamente entrenada es muy bajo. Aunque el entrenamiento de una red puede tomar mucho tiempo, este proceso solo hay que hacerlo, en general, una vez, disponiendo tras de ello de una red entrenada para usarla todas las veces deseadas.

Capítulo 2

Redes Neuronales

Las redes neuronales artificiales, conocidas simplemente como redes neuronales, son modelos de aprendizaje automático que pueden ser utilizados para multitud de problemas muy variados y diferentes entre si, desde abordar la clasificación de imágenes y la generación de texto hasta la aproximación numérica de funciones, que es para lo que serán empleadas en este trabajo.

Las redes neuronales están compuestas por una gran cantidad de elementos de procesamiento simples denominados *Neuronas*, interconectadas entre si. Es común hacer la analogía de que las redes neuronales artificiales son similares a la red neuronal que conforma el cerebro humano, donde las neuronas están interconectadas mediante conexiones sinápticas y se activan en función de los estímulos que reciban. No obstante, como veremos, las redes neuronales artificiales son una estructura matemática perfectamente definida, donde las entradas y salidas son funciones matemáticas (véase [12] y [35]).

2.1. Modelo de una Neurona

La neurona es la unidad elemental de procesamiento de información en una red neuronal. Las neuronas reciben valores numéricos (generalmente salidas de otras neuronas) y los procesan para obtener un nuevo valor numérico que transmiten a otras neuronas de la red.

Matemáticamente, el funcionamiento de la k-ésima neurona de una red neuronal artificial, como la que se describe en [35], aparece representado en la Figura 2.1, donde se pueden identificar los elementos básicos que la conforman:

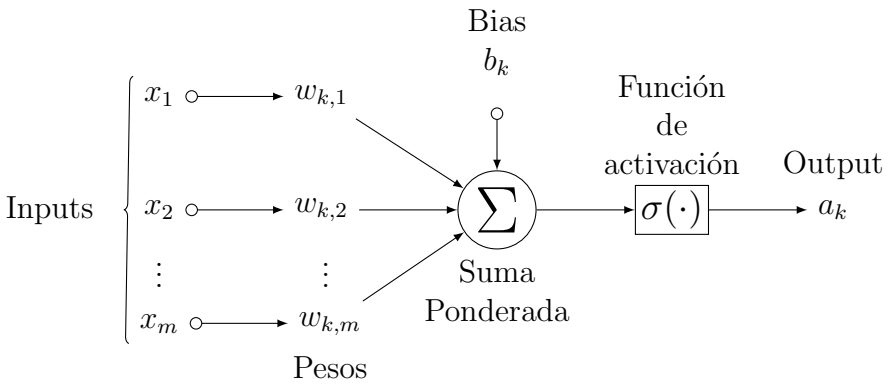


Figura 2.1: Diagrama de funcionamiento de la k-ésima neurona de una Red.

Inputs:

Las entradas $x_1, x_2, \ldots, x_m$ son valores reales que pueden ser unos datos iniciales o bien provenir de otras neuronas.

Pesos:

Los pesos $w_{k,j}$ representan la intensidad en la entrada que tiene el input j-ésimo en la neurona k-ésima de la red. A priori, los pesos pueden ser tanto positivos como negativos, es decir, $w_{k,j} \in \mathbb{R}$.

Sumatorio Ponderado:

Se calcula la combinación lineal de las señales de los inputs ponderadas por sus respectivos pesos $\sum w_{k,j} x_j$.

Bias:

El bias o sesgo b_k es una constante que se añade a la suma ponderada de los imputs para aumentar o disminuir, dependiendo de su signo, el argumento que se le pasa a la función de activación para modificar la salida de la neurona.

Función de Activación:

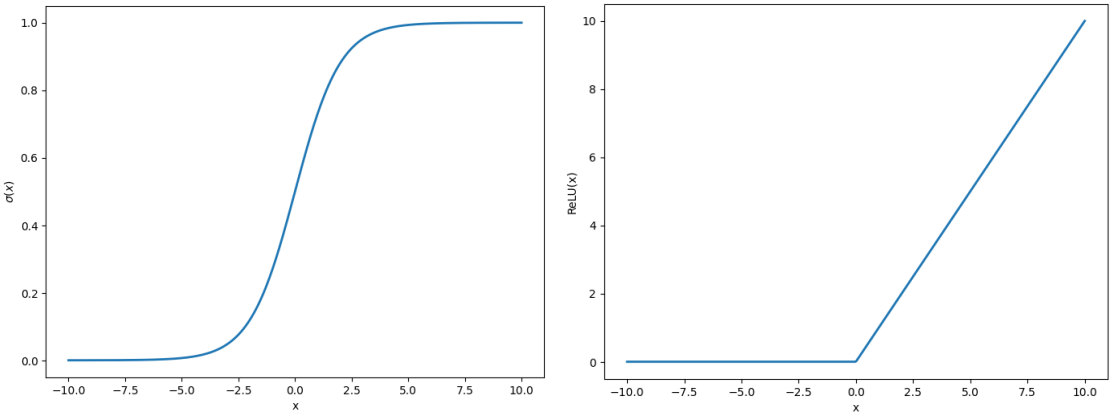
Es una función $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ (típicamente no lineal para posibilitar así la modelización de datos más complejos), que determina la salida que produce una neurona. Hay muchas clases de funciones de activación (ver [34]). La elección de unas u otras puede depender tanto de sus propiedades como de su eficacia para tratar el problema con el que se esté trabajando.

Las funciones de activación más comunes las podemos ver representadas en la Figura 2.2, donde a la izquierda aparece la función *sigmoide* o *logística* definida por

$$\sigma(x) = \frac{1}{1 + e^{-x}}, x \in \mathbb{R}\,, \tag{2.1}$$

y a la derecha la función ReLU (Rectified Linear Unit) definida por

$$\sigma(x) = \text{máx}(0, x), x \in \mathbb{R}\,. \tag{2.2}$$



(a) Función Sigmoide: $\sigma(x) = \frac{1}{1+e^{-x}}$ (b) ReLU: $\sigma(x) = \text{máx}(0, x)$

Figura 2.2: Funciones de activación más comunes.

Como se observará en las siguientes secciones, la derivada de la función de activación de una red se calcula repetidamente durante el proceso de entrenamiento. Por lo tanto, es importante elegir una función de activación cuya derivada sea fácil de computar. En el caso de la función sigmoide se puede comprobar que su derivada es

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x)), x \in \mathbb{R}\,. \tag{2.3}$$

Por lo tanto, una vez calculado el valor de la función sigmoide en un punto, se tiene automáticamente el valor de su derivada sin necesidad de realizar otra costosas evaluaciones.

Para la función ReLU, aunque no es derivable en $x = 0$, se puede establecer que su derivada sea la función de Heaviside

$$\sigma'(x) = H(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}, x \in \mathbb{R}\,. \tag{2.4}$$

La expresión matemática de la salida de la neurona k-ésima de una red es:

$$a_k = \sigma\Big(\sum_{j=1}^{m} w_{k,j} x_j + b_k\Big)\,, \tag{2.5}$$

cantidad que se denomina *activación* de la neurona.

2.2. El Perceptrón Multicapa

En esta sección se va a describir una arquitectura de red neuronal que se conoce como Perceptrón Multicapa (*Multi-layer Perceptron*). En este tipo de redes se tiene un conjunto de neuronas como las descritas en la sección anterior que se disponen en una o varias capas de neuronas interconectadas.

Supongamos que una red cuenta con L capas, donde la primera y la L-ésima capa se denominan *Capa de Input* y *Capa de Output* respectivamente. Las capas intermedias se denominan *Capas Ocultas*.

Para $l = 1, 2 \ldots, L$, sea n_l el número de neuronas de la capa l. En particular, n_1 es la dimensión de los datos de entrada y n_L es la dimensión de los datos que se producen como salida de la red.

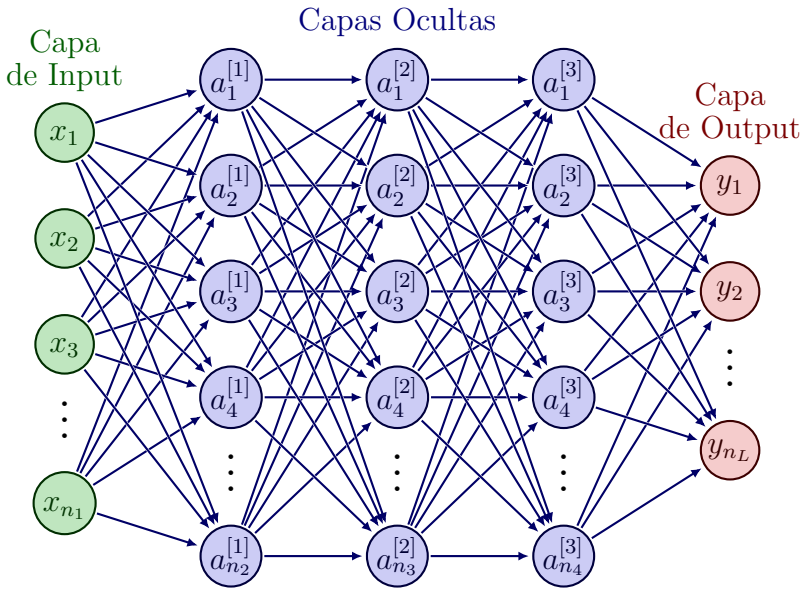


Figura 2.3: Ejemplo de un Perceptrón Multicapa con 3 capas ocultas ($L = 5$).

En general, aunque existen otras arquitecturas de redes neuronales, la Figura 2.3 muestra como las neuronas de una capa únicamente trasmiten sus salidas a las neuronas de la siguiente capa.

El objetivo es obtener una expresión recurrente para la salida de las neuronas de cada capa y finalmente, tener una expresión para la salida que produce la red

dado un input $\mathbf{x} \in \mathbb{R}^{n_1}$. Para ello se sigue el desarrollo presentado en [20].

Para $l = 1, \ldots, L$ y $k = 1, \ldots, n_l$, sea $a_k^{[l]}$ la *salida o activación* de la neurona k-ésima de la l-ésima capa.

Dada una entrada para la red $\mathbf{x} = (x_1, \ldots, x_{n_1})^T \in \mathbb{R}^{n_1}$ y una función de activación $\sigma\colon \mathbb{R} \longrightarrow \mathbb{R}$, se considera por definición que las activaciones de las neuronas de la primera capa se corresponden con el input, esto es:

$$a_k^{[1]} = x_k \text{ para } k = 1, \ldots, n_1\,. \tag{2.6}$$

Para $l = 2, \ldots, L$ y conociendo las activaciones de las neuronas de la capa $l-1$, $a_1^{[l-1]}, \ldots, a_{n_{l-1}}^{[l-1]}$, se puede generalizar la fórmula de (2.5) para definir la activación de la neurona k-ésima de la capa l como:

$$a_k^{[l]} = \sigma\Big(\sum_{j=1}^{n_{l-1}} w_{k,j}^{[l]} a_j^{[l-1]} + b_k^{[l]}\Big) \text{ para } k = 1, \ldots, n_l\,, \tag{2.7}$$

donde $w_{k,j}^{[l]}$ el *peso* que existe entre la neurona j-ésima de la capa $l-1$ y la neurona k-ésima de la capa l y $b_k^{[l]}$ es el término de *bias o sesgo* que introduce la neurona k-ésima de la capa l para $k = 1, \ldots, n_l$ y $j = 1, \ldots, n_{l-1}$.

Por comodidad, se emplea una notación matricial y vectorial para agrupar los pesos y bias, de manera que se denota

$$W^{[l]} = \Big(w_{i,j}^{[l]}\Big)_{\substack{i=1,\ldots,n_l \\ j=1,\ldots,n_{l-1}}} \in \mathbb{R}^{n_l \times n_{l-1}} \text{ para } l = 2, \ldots, L\,,$$

a la matriz de pesos de la capa l y

$$\mathbf{b}^{[l]} = (b_1^{[l]} \ldots, b_{n_l}^{[l]})^T \in \mathbb{R}^{n_l} \text{ para } l = 2, \ldots, L\,,$$

al vector de bias de la capa l.

Para trabajar con la notación que se acaba de introducir, dada una función de activación $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$ y $\mathbf{x} = (x_1, \ldots, x_m)^T \in \mathbb{R}^m$ un vector arbitrario de cualquier dimensión, se define la función de activación vectorizada $\boldsymbol{\sigma} : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ aplicando σ componente a componente. Esto es, poniendo

$$(\boldsymbol{\sigma}(\mathbf{x}))_i = \sigma(x_i) \text{ para } i = 1, \ldots, m\,.$$

Entonces, si se denota por

$$\mathbf{a}^{[l]} = (a_1^{[l]}, \ldots, a_{n_l}^{[l]})^T \in \mathbb{R}^{n_l} \text{ para } l = 1, \ldots, L\,,$$

al vector cuyas componentes son las activaciones de todas las neuronas de la l-ésima capa, se obtiene el siguiente resultado:

Proposición 2.1. *Dado* $\mathbf{x} \in \mathbb{R}^{n_1}$*, se tiene que:*

$$\mathbf{a}^{[1]} = \mathbf{x} \in \mathbb{R}^{n_1}\,. \tag{2.8}$$

$$\mathbf{a}^{[l]} = \boldsymbol{\sigma}\Big(W^{[l]}\mathbf{a}^{[l-1]} + \mathbf{b}^{[l]}\Big) \in \mathbb{R}^{n_l} \text{ para } l = 2, \ldots, L\,. \tag{2.9}$$

Este resultado proporciona un algoritmo para calcular de manera recurrente la salida de la red $\mathbf{a}^{[L]}$ para una entrada dada de la red $\mathbf{x} \in \mathbb{R}^{n_1}$.

Además, la Proposición 2.1 evidencia que, fijada una función de activación, la red neuronal establece una aplicación

$$\begin{array}{rccc} Y: & \mathbb{R}^{n_1} & \longrightarrow & \mathbb{R}^{n_L} \\ & \mathbf{x} & \longmapsto & Y(\mathbf{x}) = \mathbf{a}^{[L]} \end{array}$$

en función de todos los pesos y bias almacenados en $W^{[l]}$ y $\mathbf{b}^{[l]}$ para $l = 2, \ldots, L$.

2.3. Entrenamiento de una Red Neuronal.

Como se ha visto en la sección anterior, una red neuronal es una aplicación $Y : \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_L}$ que depende de los valores almacenados en $[\mathbf{W}, \mathbf{b}]$, los parámetros de la red, donde

$$\mathbf{W} = [W^{[2]}, W^{[3]}, \ldots, W^{[L]}] \text{ y } \mathbf{b} = [\mathbf{b}^{[2]}, \mathbf{b}^{[3]}, \ldots, \mathbf{b}^{[L]}]\,. \tag{2.10}$$

Entonces, el número de parámetros que se tienen para una red como la que se acaba de describir es

$$P = \sum_{l=2}^{L}(n_l \times n_{l-1}) + n_l = \sum_{l=2}^{L} n_l(n_{l-1} + 1)\,. \tag{2.11}$$

Por lo tanto, se puede entender que el conjunto de todos los parámetros de la red está en el espacio $\Theta = \mathbb{R}^P$.

Como se muestra en [27], esencialmente, una red neuronal es una interconexión de neuronas simples organizadas en capas que se puede describir como una aplicación

$$\mathcal{Y} : \mathbb{R}^{n_1} \times \Theta \longrightarrow \mathbb{R}^{n_L}\,, \tag{2.12}$$

descrita así para indicar explícitamente la dependencia respecto de los parámetros que tienen las salidas que produce la red.

Esta aplicación toma como inputs un vector $\mathbf{x} \in \mathbb{R}^{n_1}$ y un vector de parámetros $\theta \in \Theta$ y devuelve como output un vector $\mathcal{Y}(\mathbf{x}, \theta) \in \mathbb{R}^{n_L}$.

Sea ahora $F\colon \mathbb{R}^{n_1} \longrightarrow \mathbb{R}^{n_L}$ una función que se desea aproximar por una red similar a la descrita en secciones anteriores.

Para ello, lo primero que se precisa es una *muestra de entrenamiento*

$$\mathcal{T} = \{(\mathbf{x}_i, F(\mathbf{x}_i))\}_{i=1}^{N} \tag{2.13}$$

que es un conjunto formado por una serie de puntos del dominio de F, $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^{n_1}$ para los cuales se conocen $\{F(\mathbf{x}_i)\}_{i=1}^{N} \subset \mathbb{R}^{n_L}$, los valores de la función F en esos puntos.

El problema para hacer que la red neuronal se aproxime a la función F con los datos que se tienen en $\mathcal{T}$, consiste en encontrar la configuración de parámetros (pesos y bias) que hagan que las salidas de la red difieran lo menos posible con las salidas de la función F para los puntos de la muestra de entrenamiento.

Para tratar este problema, se define una función objetivo que se intentará minimizar:

Definición 2.2. La función $C : \Theta \longrightarrow \mathbb{R}$ definida para cada $\theta = [W, \mathbf{b}] \in \Theta$ como el error cuadrático medio

$$C(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} ||F(\mathbf{x}_i) - \mathcal{Y}(\mathbf{x}_i, \theta)||_2^2 \,. \tag{2.14}$$

se denomina *Función de Coste*.

Es razonable suponer que para que la red se aproxime adecuadamente a la función F en puntos no incluidos en la muestra de entrenamiento, se requiera que las que las salidas que produce la red neuronal $\mathcal{Y}(\mathbf{x}_i, \theta)$ para los puntos de la muestra de entrenamiento sean iguales o lo más cercanas posibles (para la norma euclídea de $\mathbb{R}^{n_L}$) a las salidas conocidas de la función $F(\mathbf{x}_i)$.

Por tanto, es lógico querer encontrar

$$\tilde{\theta} = \operatorname*{argmin}_{\theta \in \Theta} C(\theta) \,,$$

el conjunto de pesos y bias de la red que minimicen la función de coste definida en (2.14), ya que de esta manera se tiene que la distancia promedio entre los valores conocidos de los puntos $F(\mathbf{x}_i)$ y las salidas de la red $\mathcal{Y}(\mathbf{x}_i, \tilde{\theta})$ es lo menor posible.

Definición 2.3. Al proceso de buscar los pesos y sesgos que minimizan la función de coste se denomina *entrenamiento* (*train*) de la red neuronal.

2.3.1. El Descenso de Gradiente Estocástico.

En esta subsección, se describirá un algoritmo iterativo de optimización conocido como el método del descenso del gradiente, que posibilita encontrar un mínimo local de la función de coste $C : \Theta \longrightarrow \mathbb{R}$. Posteriormente, se examinará una variante de este algoritmo que reduce el coste computacional de manera significativa para su implementación en la práctica, denominada descenso de gradiente estocástico.

El algoritmo del descenso del gradiente se basa en el recurso de que dada una función real $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, para cada $\mathbf{x}_0 \in \mathbb{R}^n$, el vector gradiente de f en $\mathbf{x}_0$

$$\nabla f(\mathbf{x}_0) = \Big[\frac{\partial f}{\partial x_1}(\mathbf{x}_0), \frac{\partial f}{\partial x_2}(\mathbf{x}_0), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x}_0)\Big]^T ,$$

marca la dirección de variación máxima de la función en $\mathbf{x}_0$, por tanto, si se desea de minimizar la función f, hay que moverse en la dirección opuesta a la que indica el vector gradiente.

Se empleará esta técnica con el objetivo de minimizar la función de coste calculando el gradiente de la función de coste, es decir, $\nabla C(\theta)$.

Para ello, asumiendo que la función de coste C es diferenciable y empleando la fórmula de Taylor para funciones de varias variables (véase [15]). tenemos para $\theta \in \Theta$, una aproximación local de la función C en un parámetro arbitrario de la red:

$$C(\theta + \Delta\theta) = C(\theta) + \nabla C(\theta)(\Delta\theta) + \varepsilon(\Delta\theta)||\Delta\theta||^2 ,$$

para $\Delta\theta \in \Theta$ un vector tal que $\theta + \Delta\theta$ sea suficientemente próximo a θ y ε una funcion tal que

$$\lim_{\Delta\theta \to \mathbf{0}} \varepsilon(\Delta\theta) = 0 .$$

Por lo tanto, si $\Delta\theta \in \Theta$ es suficientemente próximo a $\mathbf{0} \in \Theta$ se puede despreciar el término $||\Delta\theta||^2$. Teniendo en cuenta que

$$\nabla C(\theta)(\Delta\theta) = \langle \nabla C(\theta), \Delta\theta \rangle ,$$

se puede considerar que

$$C(\theta + \Delta\theta) \approx C(\theta) + \nabla C(\theta)^T \Delta\theta \in \mathbb{R} \tag{2.15}$$

para $\Delta\theta \in \Theta$ suficientemente próximo a $\mathbf{0} \in \Theta$.

Por lo tanto, dado que se desea minimizar la función de coste, se intentará encontrar un vector $\Delta\theta$ de manera que la cantidad $\nabla C(\theta)^T \Delta\theta$ sea lo más pequeña posible.

Para ello se emplea la desigualdad de Cauchy-Schwarz, que sostiene que para todos $\mathbf{u}, \mathbf{v} \in \Theta$ se cumple que

$$|\langle \mathbf{u}, \mathbf{v} \rangle| = |\mathbf{u}^T \mathbf{v}| \leq ||\mathbf{u}||_2 ||\mathbf{v}||_2 \,. \tag{2.16}$$

Además, es un hecho conocido que la igualdad en (2.16) solo se alcanza si y solamente si $\mathbf{u}$ y $\mathbf{v}$ son linealmente dependientes.

Con lo cual, $\nabla C(\theta)^T \Delta\theta$ solo puede ser tan pequeño como $-||\nabla C(\theta)||_2 ||\Delta\theta||_2$, cosa que solo sucede cuando $\Delta\theta$ esta en la dirección de $\nabla C(\theta)$, o lo que es lo mismo, cuando

$$\Delta\theta = -\nabla C(\theta) \,. \tag{2.17}$$

Teniéndose en cuenta que la aproximación de (2.15) solo vale para $\Delta\theta$ próximo a $\mathbf{0}$, si se fija un $\eta \in \mathbb{R}$ suficientemente pequeño, que se conce como *Tasa de Aprendizaje*, se tiene que el vector $\theta - \eta \nabla C(\theta)$ da un pequeño paso en la dirección que marca (2.17) hacia un mínimo local.

Entonces, empleando los argumentos que se acaban de describir, se puede construir un algoritmo iterativo que converja hacia un mínimo local de la función de coste.

Para su implementación práctica, el procedimiento debe poseer algún criterio de parada, como el establecimiento de una tolerancia o de un número máximo de iteraciones, para no exceder la capacidad computacional de la que se disponga.

El algoritmo que se proporciona es el siguiente:

Datos: Función objetivo C, vector inicial $\theta \in \Theta$, tolerancia tol y $maxiter$.
$n \leftarrow 1$
mientras $C(\theta) > tol$ y $n \leq maxiter$ **hacer**
  $\theta \leftarrow \theta - \eta \nabla C(\theta)$
  $n \leftarrow n + 1$
fin

Algoritmo 1: Descenso del Gradiente

Este algoritmo presenta un inconveniente. Para cada punto $\mathbf{x}_i$ de la muestra de entrenamiento (2.13), se denota

$$C_{\mathbf{x}_i}(\theta) = \frac{1}{2} ||F(\mathbf{x}_i) - \mathcal{Y}(\mathbf{x}_i, \theta)||_2^2 \,. \tag{2.18}$$

En cada iteración se precisa calcular el gradiente de la función de coste, es

decir,

$$\nabla C(\theta) = \nabla\Big(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}||F(\mathbf{x}_i) - \mathcal{Y}(\mathbf{x}_i,\theta)||_2^2\Big) = \frac{1}{N}\sum_{i=1}^{N}\nabla C_{\mathbf{x}_i}(\theta)\,,$$

lo que muestra que calcular el gradiente de la función de coste equivale a tener que calcular una suma de gradientes que recorre toda la muestra de entrenamiento.

Por lo tanto, este hecho supone que si N, el tamaño de la muestra de entrenamiento, es grande, el coste de calcular $\nabla C(\theta)$ para cada iteración puede tener llegar a tener un coste computacional inasumible.

Para evitar tener que calcular todos los gradientes de todos los datos de la muestra de entrenamiento en cada iteración, se introduce una variante al Algoritmo 1 denominada descenso del gradiente estocástico. Este método, consiste en escoger de forma aleatoria un subconjunto de los datos de (2.13), y realizar una iteración del descenso del gradiente únicamente con los datos seleccionados, aunque la versión del algoritmo que se dará, considera solamente un único dato en cada iteración del descenso del gradiente.

Datos: Vector inicial $\theta \in \Theta$, tolerancia *tol* y *maxiter*.
$n \leftarrow 1$
mientras $C(\theta) > tol$ *y* $n \leq maxiter$ **hacer**
  $i \leftarrow$ entero aleatorio elegido de $\{1,\ldots,N\}$
  $\theta \leftarrow \theta - \eta\nabla C_{\mathbf{x}_i}(\theta)$
  $n \leftarrow n+1$
fin

Algoritmo 2: Descenso del Gradiente Estocástico.

Se observa que a medida que se van realizando las sucesivas iteraciones, se van empleando diferentes datos de los que se calcula su respectivo gradiente. Conviene advertir que aunque el coste computacional de cada iteración se reduce enormemente, no está garantizado que se reduzca la función de coste, empíricamente es un método que ha probado su utilidad en muchos modelos (ver [29]) y el grandísimo ahorro computacional que se obtiene a cambio hace que merezca la pena ser probado.

En función de cómo se escojan los puntos de la muestra de entrenamiento (con remplazamiento, sin remplazamiento, etc.) se pueden dar multitud de variantes del algoritmo del descenso del gradiente estocástico que aparecen recogidas con detalle en [4].

2.3.2. El Algoritmo de Backpropagation.

En la subsección anterior se ha expuesto el proceso mediante el cual las redes neuronales aprenden empleando el descenso del gradiente estocástico. Para aplicar ese algoritmo es necesario computar muchas veces el gradiente de la función de coste de cada dato de entrada, es decir, $\nabla C_{\mathbf{x}_i}$.

Para calcular dicho gradiente, es necesario ver cómo de sensible es la función de coste (2.18) respecto de cada parámetro, que son todos los pesos $w_{k,j}^{[l]}$ y bias $b_j^{[l]}$.

Tomando $F(\mathbf{x}_i) = \mathbf{y}$, se reescribe (2.18) como

$$C = \frac{1}{2}||\mathbf{y} - \mathbf{a}^{[L]}||_2^2\,, \tag{2.19}$$

para obviar la dependencia del dato $\mathbf{x}_i$.

El propósito de esta subsección es el de porporcionar un método para calcular las derivadas parciales de la función de coste respecto de cada parámetro

$$\frac{\partial C}{\partial w_{j,k}^{[l]}}\,, \frac{\partial C}{\partial b_j^{[l]}} \text{ para } l = 2, \ldots, L\,, \tag{2.20}$$

de manera sencilla con la finalidad de aplicar el método del gradiente estocástico de manera eficiente.

Observación 3. Es conveniente notar que la dependencia de C respecto de los parámetros recae exclusivamente en el output $\mathbf{a}^{[\mathbf{L}]}$ que produce la red como salida.

Una posible estrategia abordar el problema de calcular las derivadas de (2.20) es aproximarlas utilizando diferencias finitas. Al ordenar todos los parámetros de la red (pesos y bias) en el vector $\theta \in \Theta$, para un valor de h suficientemente pequeño se tiene que

$$\frac{\partial C}{\partial \theta_j} = \frac{C(\theta + h\mathbf{e}_j) - C(\theta)}{h}\,,$$

donde $\mathbf{e}_j$ es el j-ésimo vector canónico de $\Theta = \mathbb{R}^P$.

Este enfoque es sencillo a nivel conceptual y fácil de implementar en un ordenador, pero presenta una grave deficiencia. Si la red tiene un gran número de parámetros, el cálculo de las derivadas parciales requiere evaluar la función de coste tantas veces como parámetros existan. Además, evaluar la función de coste requiere realizar un paso hacia adelante sobre toda la red (*forward pass*), lo que resulta en un coste computacional muy elevado.

Para superar estos inconvenientes, se introduce un algoritmo conocido como *Backpropagation*. Este término se refiere al hecho de que, para calcular el gradiente

de la función de coste, el algoritmo de *Backpropagation* requiere únicamente un paso hacia adelante y un paso hacia atrás (*backward pass*), con el mismo coste computacional que el paso hacia adelante. Además, este algoritmo calcula todas las derivadas de (2.20) de manera simultánea, en lugar de hacerlo individualmente como en el método anterior.

Como se expone en [30], el algoritmo de *Backpropagation* fue desarrollado inicialmente en la década de 1970. Sin embargo, no fue hasta 1986 cuando David Rumelhart, Geoffrey Hinton y Ronald Williams señalaron por primera vez en [33] como la *Backpropagation* funciona de manera más rápida que otros los enfoques que se habían intentado hasta entonces en el campo del Aprendizaje Automático. Este hecho propició un auge en la investigación de las redes neuronales, ya que con la reducción del coste computacional, se aumentó la gama de problemas que las redes neuronales podían afrontar.

Para calcular las derivadas parciales de (2.20) empleando la *Backpropagation* se introducirán dos nuevas cantidades auxiliares que serán de gran utilidad durante el desarrollo de esta sección.

Definición 2.4. Se define la *Entrada Ponderada* de la j-ésima neurona de la l-ésima capa como la j-ésima componente del vector

$$\mathbf{z}^{[l]} = [z_1^{[l]}, \ldots, z_{n_l}^{[l]}]^T = W^{[l]}\mathbf{a}^{[l-1]} + \mathbf{b}^{[l]} \in \mathbb{R}^{n_l} \text{ para } l = 2, \ldots, L\,.$$

Observación 4. Es de inmediato comprobar que para la función de activación vectorizada $\boldsymbol{\sigma}$ que se considere, se mantiene la relación

$$\mathbf{a}^{[l]} = \boldsymbol{\sigma}\left(\mathbf{z}^{[l]}\right) \text{ para } l = 2, \ldots, L \tag{2.21}$$

Definición 2.5. Se define el *Error* de la j-ésima neurona de la l-ésima capa como

$$\delta_j^{[l]} = \frac{\partial C}{\partial z_j^{[l]}} \text{ para } j = 1, \ldots, n_l \text{ y } l = 2, \ldots, L \tag{2.22}$$

que se dispone por componentes en el vector $\boldsymbol{\delta}^{[l]} \in \mathbb{R}^{n_l}$.

También, es necesario definir una nueva operación entre vectores.

Definición 2.6. Se define el *Producto de Hadamard* de dos vectores $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ como un nuevo vector $\mathbf{u} \odot \mathbf{v} \in \mathbb{R}^n$ donde sus componentes vienen dadas por

$$(\mathbf{u} \odot \mathbf{v})_i = u_i v_i \text{ para todo } i = 1, \ldots, n \tag{2.23}$$

Teniendo en cuenta todas estas consideraciones preliminares, se puede dar el siguiente resultado cuya demostración tomamos de [20] y [30].

Proposición 2.7 (Ecuaciones Fundamentales de la Backpropagation)**.** *Para una función de activación* $\boldsymbol{\sigma}$ *se verifica que*

$$\boldsymbol{\delta}^{[L]} = \boldsymbol{\sigma}'(\mathbf{z}^{[L]}) \odot (\mathbf{a}^{[L]} - \mathbf{y}) \tag{2.24}$$

$$\boldsymbol{\delta}^{[l]} = \boldsymbol{\sigma}'(\mathbf{z}^{[l]}) \odot \left(W^{[l+1]}\right)^T \boldsymbol{\delta}^{[l+1]} \qquad \text{para } l = 2, \ldots, L-1 \tag{2.25}$$

$$\frac{\partial C}{\partial b_j^{[l]}} = \delta_j^{[l]} \qquad \text{para } l = 2, \ldots, L \tag{2.26}$$

$$\frac{\partial C}{\partial w_{j,k}^{[l]}} = \delta_j^{[l]} a_k^{[l-1]} \qquad \text{para } l = 2, \ldots, L \tag{2.27}$$

Demostración. Primero se prueba la relación (2.24).

Para cada $j = 1, \ldots, n_L$ se debe determinar la expresión de $\delta_j^{[L]} = \frac{\partial C}{\partial z_j^{[L]}}$. Esto se logra combinando las identidades (2.19) y (2.21), obteniendo que:

$$C = \frac{1}{2}||\mathbf{y} - \mathbf{a}^{[L]}||_2^2 = \frac{1}{2}\sum_{k=1}^{n_L}(y_k - a_k^{[L]})^2 = \frac{1}{2}\sum_{k=1}^{n_L}(y_k - \sigma(z_k^{[L]}))^2 \,. \tag{2.28}$$

Al aplicar la versión de la regla de la cadena que se propone en [15] a la expresión previa y derivar con respecto a $z_j^{[L]}$, se obtiene que:

$$\delta_j^{[L]} = \frac{\partial C}{\partial z_j^{[L]}} = \frac{\partial C}{\partial a_j^{[L]}} \frac{\partial a_j^{[L]}}{\partial z_j^{[L]}} \text{ para cada } j = 1 \ldots, n_L \,.$$

Los componentes de esta expresión son fáciles de calcular, puesto que $a_j^{[L]} = \sigma(z_j^{[L]})$, implica directamente que:

$$\frac{\partial a_j^{[L]}}{\partial z_j^{[L]}} = \sigma'(z_j^{[L]}) \,.$$

Además, al derivar (2.28) con respecto a $a_j^{[L]}$ es claro que:

$$\frac{\partial C}{\partial a_j^{[L]}} = -(y_j - a_j^{[L]}) \,.$$

Entonces, juntando los dos términos recién calculados se tiene que:

$$\delta_j^{[L]} = \frac{\partial C}{\partial z_j^{[L]}} = (a_j^{[L]} - y_j)\sigma'(z_j^{[L]}) \text{ para cada } j = 1, \ldots, n_L$$

lo cual, expresado en notación vectorial con el producto de Hadamard resulta en:

$$\boldsymbol{\delta}^{[L]} = \boldsymbol{\sigma}'(\mathbf{z}^{[L]}) \odot (\mathbf{a}^{[L]} - \mathbf{y}) \,.$$

Para demostrar (2.25) se expresarán los errores de la capa l en función de los errores de la capa $l+1$, ya que la variación de la función de coste respecto a la variación de la entrada ponderada $z_j^{[l]}$ dependerá de las variaciones introducidas por todas las neuronas de la capa posterior en función de las entradas que reciben de la capa anterior.

Entonces, a partir de la definición de error y aplicando la regla de la cadena se tiene que:

$$\delta_j^{[l]} = \frac{\partial C}{\partial z_j^{[l]}} = \sum_{k=1}^{n_{l+1}} \frac{\partial C}{\partial z_k^{[l+1]}} \frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = \sum_{k=1}^{n_{l+1}} \delta_k^{[l+1]} \frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} \text{ para cada } j = 1, \ldots, n_l \quad (2.29)$$

Por construcción de las entradas ponderadas se tiene que:

$$\mathbf{z}^{[l+1]} = W^{[l+1]}\mathbf{a}^{[l]} + \mathbf{b}^{[l+1]} \in \mathbb{R}^{n_{l+1}} \,.$$

luego, las componentes del vector anterior se escriben como:

$$z_k^{[l+1]} = \sum_{s=1}^{n_l} w_{k,s}^{[l+1]} \sigma(z_s^{[l]}) + b_k^{[l+1]} \text{ para cada } k = 1, \ldots, n_{l+1} \,.$$

Entonces, derivando la expresión anterior con respecto a las entradas ponderadas de la capa anterior se obtiene que:

$$\frac{\partial z_k^{[l+1]}}{\partial z_j^{[l]}} = w_{k,j}^{[l+1]} \sigma'(z_j^{[l]}) \text{ para cada } k = 1, \ldots, n_{l+1} \text{ y } l = 1, \ldots, n_l \,. \quad (2.30)$$

Juntando las expresiones (2.29) y (2.30) se llega a que:

$$\delta_j^{[l]} = \sum_{k=1}^{n_{l+1}} \delta_k^{[l+1]} w_{k,j} \sigma'(z_j^{[l]}) \text{ para cada } j = 1, \ldots, n_l$$

lo que reescrito en notación matricial y vectorial queda de la forma

$$\delta_j^{[l]} = \sigma'(z_j^{[l]})\Big(\Big(W^{[l+1]}\Big)^T \boldsymbol{\delta}^{[l+1]}\Big)_j \text{ para cada } j = 1, \ldots, n_l\,.$$

Estas son las componentes del producto de Hadamard

$$\boldsymbol{\sigma}'(\mathbf{z}^{[l]}) \odot \Big(\Big(W^{[l+1]}\Big)^T \boldsymbol{\delta}^{[l+1]}\Big)\,,$$

lo que prueba la relación (2.25).

Para demostrar (2.26) para cada $l = 2, \ldots, L$, se observa que la función de coste depende de $z_j^{[l]}$ para cada $j = 1, \ldots, n_l$ y que por construcción, evidentemente $z_j^{[l]}$ depende del bias $b_j^{[l]}$. Por lo tanto, en virtud de la regla de la cadena se tiene que:

$$\frac{\partial C}{\partial b_j^{[l]}} = \frac{\partial C}{\partial z_j^{[l]}} \frac{\partial z_j^{[l]}}{\partial b_j^{[l]}} \text{ para cada } j = 1, \ldots, n_l\,. \tag{2.31}$$

El primer factor ya se conoce, pues es $\delta_j^{[l]}$, por lo que resta calcular $\frac{\partial z_j^{[l]}}{\partial b_j^{[l]}}$. Para ello, se observa que:

$$z_j^{[l]} = \Big(\sum_{k=1}^{n_{l-1}} w_{j,k}^{[l]} \sigma(z_k^{[l-1]})\Big) + b_j^{[l]} \text{ para cada } j = 1, \ldots, n_l\,, \tag{2.32}$$

y dado que $z_k^{[l-1]}$ no depende $b_j^{[l]}$, es claro que:

$$\frac{\partial z_j^{[l]}}{\partial b_j^{[l]}} = 1\,.$$

Luego, en virtud de (2.31), queda demostrado que:

$$\frac{\partial C}{\partial b_j^{[l]}} = \delta_j^{[l]} \text{ para cada } j = 1, \ldots, n_l\,.$$

Finalmente, para probar (2.27), si para cada $l = 2, \ldots, L$ se deriva la función de coste con respecto a cada elemento de $W^{[l]}$ aplicando la regla de la cadena, se obtiene:

$$\frac{\partial C}{\partial w_{j,k}^{[l]}} = \sum_{s=1}^{n_l} \frac{\partial C}{\partial z_s^{[l]}} \frac{\partial z_s^{[l]}}{\partial w_{j,k}^{[l]}}\,. \tag{2.33}$$

De la expresión de (2.32) se deduce que, independientemente del valor de $j = 1, \ldots, n_l$ se tiene que:

$$\frac{\partial z_j^{[l]}}{\partial w_{j,k}^{[l]}} = a_k^{[l-1]} \text{ para todo } k = 1, \ldots, n_{l-1}\,.$$

Además, puesto que la entrada ponderada $z_s^{[l]}$ solo depende de los pesos de las capas anteriores y de $w_{s,k}^{[l]}$ para $k = 1, \ldots, n_{l-1}$, se deduce que:

$$\frac{\partial z_s^{[l]}}{\partial w_{j,k}^{[l]}} = 0 \text{ si } s \neq j\,.$$

Entonces, aplicando estas dos observaciones en (2.33) y usando la definición de error de una capa, se llega a que:

$$\frac{\partial C}{\partial w_{j,k}^{[l]}} = \sum_{s=1}^{n_l} \frac{\partial C}{\partial z_s^{[l]}} \frac{\partial z_s^{[l]}}{\partial w_{j,k}^{[l]}} = \frac{\partial C}{\partial z_j^{[l]}} \frac{\partial z_j^{[l]}}{\partial w_{j,k}^{[l]}} = \delta_j^{[l]} a_k^{[l-1]}\,,$$

como se quería demostrar. □

Observación 5. Las relaciones (2.24) y (2.25) muestran una forma de calcular iterativamente los errores de las neuronas capa l en función de los errores de las neuronas de la capa $l + 1$.

Conviene notar además que los errores de las capas están expresados en términos de la derivada de la función de activación que empleemos.

Por ejemplo si se emplea la función *sigmoide* o *logística*, que como se ha comprobado en (2.3) tiene como derivada a la función

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))\,,\ x \in \mathbb{R}\,,$$

que aparece representada en rojo en la Figura 2.4 junto con el grafo de la función sigmoide en azul.

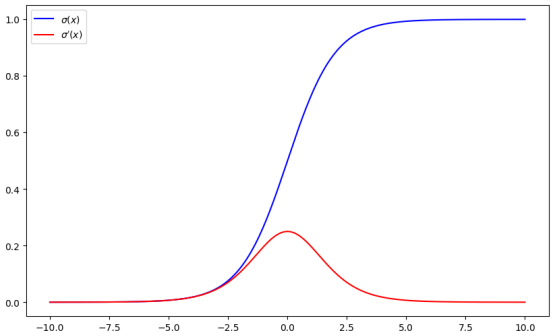


Figura 2.4: Gráfico de la función sigmoide y su derivada.

Entonces, en virtud de las relaciones de (2.26) y (2.27), se tiene que los pesos y bias de las neuronas que tengan una salida $\sigma(z_l^{[l]})$ próxima a 0 o a 1, conocidas como *neuronas saturadas*, aprenden de manera más lenta, debido a que, como se observa en la Figura 2.4:

$$\sigma'(z_j^{[l]}) \approx 0\,.$$

Este hecho en redes neuronales con muchas capas puede dar lugar al problema conocido como desvanecimiento del gradiente (*Vanishing Gradient*), ya que para calcular las derivadas parciales de los parámetros se van multiplicando sucesivamente los errores entre si. Entonces, si estos errores se vuelven muy pequeños, las actualizaciones de los pesos y bias durante el entrenamiento también se vuelven muy pequeñas, lo que provoca un aprendizaje ineficiente de la red.

Para prevenir esta interrupción en el aprendizaje, se pueden emplear otras funciones de activación, como la ReLU.

Observación 6. La relación (2.26) establece que la tasa de cambio de la función de coste con respecto a cualquier bias de la capa l de la red es igual al error de la neurona correspondiente a dicho bias, por lo que una vez computados los errores de las neuronas, se tienen automáticamente las derivadas parciales de la función de coste con respecto a los bias.

Se puede observar también que la relación (2.27) mantiene que la tasa de cambio de la función de coste respecto a un determinado peso es igual a la activación de la neurona de la que sale el peso por el error de la neurona a la que llega el peso. Entonces, puesto que se realiza un paso hacia adelante para calcular las activaciones y un paso hacia atrás en la red para computar los errores de las capas, (2.27) afirma que para calcular (2.20) no se requiere mayor coste computacional que una multiplicación extra por peso. Además, se verifica que para cada $l = 2 \ldots, L$

$$\Big(\frac{\partial C}{\partial w^{[l]}_{j,k}}\Big)_{\substack{j=1,\dots,n_l\\k=1,\dots,n_{l-1}}} = \boldsymbol{\delta}^{[l]}(\mathbf{a}^{[l-1]})^T \in \mathbb{R}^{n_l \times n_{l-1}}\,. \tag{2.34}$$

Anteriormente, se ha demostrado cómo calcular el gradiente de la función de coste de manera eficiente utilizando el algoritmo de *Backpropagation*. A continuación, se resumirá todo lo presentado en esta sección, proporcionando el pseudocódigo de un algoritmo que permite a una red neuronal aprender mediante una variante del descenso del gradiente estocástico.

Para evitar la notación con los productos de Hadamard y poder escribir un algoritmo de manera más compacta, para cada $l = 2, \dots, L$ se definen las matrices

$$D^{[l]} = \Big(d^{[l]}_{i,j}\Big)_{\substack{i=1,\dots,n_l\\j=1,\dots,n_l}} \in \mathbb{R}^{n_l \times n_l}\,,\ \text{donde } d^{[l]}_{i,j} = \begin{cases} \sigma'(z^{[l]}_i) & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases} \tag{2.35}$$

con las que por construcción se tiene que $D^{[l]}$ es una matriz cuadrada en cuya diagonal están las componentes de $\mathbf{z}^{[l]}$ y que

$$\boldsymbol{\delta}^{[L]} = D^{[L]}(\mathbf{a}^{[L]} - \mathbf{y}) \tag{2.36}$$

$$\boldsymbol{\delta}^{[l]} = D^{[l]}\Big(W^{[l+1]}\Big)^T \boldsymbol{\delta}^{[l+1]} \qquad \text{para } l = 2, \dots, L-1 \tag{2.37}$$

lo que al desarrollarse, para cada $l = 2, \dots, L$ permite expresar el vector de los errores de las neuronas de cada capa es una sola línea:

$$\boldsymbol{\delta}^{[l]} = D^{[l]}\Big(W^{[l+1]}\Big)^T D^{[l+1]}\Big(W^{[l+2]}\Big)^T \cdots D^{[L-1]}\Big(W^{[L]}\Big)^T(\mathbf{a}^{[l]} - \mathbf{y})\,. \tag{2.38}$$

Por lo tanto, se puede presentar el siguiente algoritmo de aprendizaje para una red neuronal que devuelve las matrices de pesos y los vectores de sesgo con los parámetros óptimos con la notación introducida en (2.19).

Datos: Conjunto de entrenamiento $\mathcal{T}$, número de capas L, Pesos $W^{[l]}$, Bias $b^{[l]}$, función de activación σ, tasa de aprendizaje η y número de iteraciones *numiter*.

para *n=1,. . . , numiter* **hacer**

  $k \leftarrow$ entero aleatorio elegido de $\{1, \ldots, N\}$

  $\mathbf{a}^{[1]} \leftarrow \mathbf{x}_k$

  para $l = 2, \ldots, L$ **hacer**

    $\mathbf{z}^{[l]} \leftarrow W^{[l]}\mathbf{a}^{[l-1]} + \mathbf{b}^{[l]}$

    $\mathbf{a}^{[l]} = \boldsymbol{\sigma}(\mathbf{z}^{[l]})$

    $D^{[l]} \leftarrow diag(\boldsymbol{\sigma}'(\mathbf{z}^{[l]}))$

  fin

  $\boldsymbol{\delta}^{[L]} \leftarrow D^{[L]}(\mathbf{a}^{[L]} - \mathbf{y})$

  para $l = L - 1 \ldots, 2$ **hacer**

    $\boldsymbol{\delta}^{[l]} \leftarrow D^{[l]}\left(W^{[l+1]}\right)^T \boldsymbol{\delta}^{[l+1]}$

  fin

  para $l = L, \ldots, 2$ **hacer**

    $W^{[l]} \leftarrow W^{[l]} - \eta\boldsymbol{\delta}^{[l]}(\mathbf{a}^{[l-1]})^T$

    $\mathbf{b}^{[l]} \leftarrow \mathbf{b}^{[l]} - \eta\boldsymbol{\delta}^{[l]}$

  fin

fin

Algoritmo 3: Algoritmo de Aprendizaje de una Red Neuronal.

En la práctica, se suele utilizar una versión alternativa de este algoritmo en la cual, la muestra de entrenamiento se divide en subconjuntos de tamaño uniforme denominados *lotes* o *batches*, los cuales se emplean en cada iteración del descenso del gradiente. Generalmente, durante el entrenamiento, la red recorre todos los datos de la muestra de entrenamiento varias veces. A completar el procesamiento de todo el conjunto de datos de entrenamiento una vez se le conoce como recorrer una *época*.

2.4. La Red Neuronal como Aproximador Universal.

En las secciones anteriores se ha descrito un método para aproximar una función mediante una red neuronal, sin embargo, no tenemos ninguna garantía matemática de que esta aproximación sea efectiva. Es natural preguntarse entonces sobre qué funciones es capaz de aproximar una red neuronal y cómo de buena puede ser esta aproximación.

Estas cuestiones fueron resueltas en 1989 por George V. Cybenko, quien en

[13], demostró que cualquier función continua definida en el cubo unidad, puede ser aproximada, con cualquier grado de precisión, por una red neuronal de una sola capa oculta con un número suficiente de neuronas empleando una función de activación sigmoidal.

Después de la publicación de este artículo, se desarrollaron numerosos trabajos que ampliaban los resultados obtenidos por Cybenko. Uno de los resultados de aproximación más significativos es el presentado en [23]. En este trabajo se demuestra que una red con una función de activación continua tiene capacidad de aproximación universal si y solamente si la función de activación no es polinómica.

En esta sección se expondrá el trabajo realizado por Cybenko, tanto por su relevancia histórica como por la brevedad de las demostraciones de los resultados obtenidos. Para fundamentar estos resultado, es necesario presentar una serie de conceptos y definiciones, así como teoremas de Análisis Funcional, cuyas demostraciones se pueden consultar en [32].

Proposición 2.8. *Sea $(E, ||\cdot||)$ un espacio vectorial normado y sea $F \subseteq E$ un subespacio.*

Entonces $\overline{F}$ también es un subespacio de E.

Definición 2.9. Sea $(E, ||\cdot||)$ un espacio vectorial normado sobre el cuerpo $\mathbb{K}$.

Se define su *dual topológico* como:

$$E^* = \{T\colon E \longrightarrow \mathbb{K} :\ T \text{ lineal y continua}\}. \tag{2.39}$$

A los elementos de E^* se les denomina funcionales lineales.

Teorema 2.10 (Teorema de Hahn-Banach)**.** *Sea $(E, ||\cdot||)$ un espacio vectorial normado, $F \subseteq E$ un subespacio vectorial y $T \in E^*$.*

Entonces, existe un funcional lineal $\tilde{T}\colon E \longrightarrow \mathbb{K}$ tal que

1. *$\tilde{T}|_F = T$, es decir, $\tilde{T}$ es una extensión de T.*
2. *$||\tilde{T}|| = ||T||$.*

Proposición 2.11. *Sea (E, d) un $\mathbb{R}$-espacio vectorial normado, $F \subseteq E$ un subsespacio y $x_0 \in E$ con $d(x_0, F) = d > 0$.*

Entonces, existe $T\colon E \longrightarrow \mathbb{R}$ tal que $||T|| = 1$, $||T(x_0)|| = d$ y $T|_F \equiv 0$.

Teorema 2.12 (Teorema de Representación de Riesz para espacios de Hilbert.)**.** *Sea $(H, \langle\cdot,\cdot\rangle)$ un espacio de Hilbert sobre $\mathbb{R}$ y sea $f\colon H \longrightarrow \mathbb{R}$.*

Entonces existe $y \in H$ tal que para todo $x \in H$, $f(x) = \langle x, y\rangle$

Definición 2.13. Sea $I = [0,1]$ y sea $n \in \mathbb{N}$. Entonces I^n denota al cubo unidad n-dimensional.

Se pretende estudiar la aproximación de una función continua por una expresión del tipo (2.12). Para ello se considera el espacio métrico de las funciones continuas en el cubo unidad n-dimensional dotadas de la norma infinito o del supremo $(\mathcal{C}(I^n), ||\cdot||_\infty)$.

Se recuerda que un conjunto $D \subseteq (\mathcal{C}(I^n), ||\cdot||_\infty)$ es denso si $\overline{D} = \mathcal{C}(I^n)$, o equivalentemente, si para todo $\varepsilon > 0$ y para toda función $f \in \mathcal{C}(I^n)$, existe $g \in D$ tal que $||f - g||_\infty < \varepsilon$.

Definición 2.14. Se denota por $\mathcal{M}(I^n)$ al espacio de las medidas, finitas, signadas y regulares de Borel en I^n.

Se recuerda que una medida de Borel en I^n es una medida definida sobre el espacio medible $(I^n, \mathcal{B}(I^n))$, donde $\mathcal{B}(I^n)$ es la σ-álgebra de Borel del cubo unidad n-dimensional.

Además, una medida de Borel μ en I^n se dice que es *regular* si para todo $A \in \mathcal{B}(I^n)$ se tiene que:

$$\begin{cases} \mu(A) = \text{ínf}\{\mu(U)\colon A \subset U\,, U \text{ abierto}\} \\ \mu(A) = \sup\{\mu(U)\colon A \subset K\,, K \text{ compacto}\} \end{cases} . \tag{2.40}$$

Definición 2.15. Una función $\sigma\colon \mathbb{R} \longrightarrow \mathbb{R}$ se dice que es *discriminatoria* si para una medida $\mu \in \mathcal{M}(I^n)$ y para todos $\mathbf{w} \in \mathbb{R}^n$ y $b \in \mathbb{R}$ la condición

$$\int_{I^n} \sigma(\mathbf{w}^T\mathbf{x} + b) d\mu(\mathbf{x}) = 0\,, \tag{2.41}$$

implica que $\mu = 0$.

Definición 2.16. Sea $\sigma\colon \mathbb{R} \longrightarrow \mathbb{R}$. Se dice que σ es una función *sigmoidal* si

$$\sigma(t) \to \begin{cases} 1 \text{ cuando } t \to +\infty \\ 0 \text{ cuando } t \to -\infty \end{cases} . \tag{2.42}$$

Teorema 2.17. *Sea σ una función discriminatoria continua y sea*

$$D = \left\{ G(\mathbf{x}) = \sum_{j=1}^{N} \alpha_j \sigma(\mathbf{w}^T\mathbf{x} + b)\colon N \in \mathbb{N}\,, \mathbf{w} \in \mathbb{R}^n\,, \alpha_j, b \in \mathbb{R} \right\} \subseteq \mathcal{C}(I^n)\,, \tag{2.43}$$

el conjunto de todas las combinaciones lineales finitas de esa forma.

Entonces, D es denso en $(\mathcal{C}(I^n), ||\cdot||_\infty)$.

Demostración. Razonamos por reducción al absurdo.

Supongamos que D no es un subconjunto denso de $\mathcal{C}(I^n)$. Dado que D es un subespacio vectorial de $\mathcal{C}(I^n)$, en virtud de la Proposición 2.8 se tiene que $\overline{D}$ es un subespacio propio de $\mathcal{C}(I^n)$.

Consideremos $f \in \mathcal{C}(I^n) \setminus \overline{D}$. Esto supone que $d_\infty(f, \overline{D}) > 0$ ya que si $d_\infty(f, \overline{D}) = 0$, como $\overline{D}$ es cerrado se tendría que $f \in \overline{D}$, en contra de lo que se ha supuesto.

Entonces, en virtud de la Proposición 2.11, que es consecuencia del teorema de Hahn-Banach, existe un funcional lineal acotado no nulo $L\colon \mathcal{C}(I^n) \longrightarrow \mathbb{R}$ tal que $L(\overline{D}) = L(D) = \{0\}$ (i.e. $L|_{\overline{D}} = 0$).

Si consideramos en $\mathcal{C}(I^n)$ el producto escalar definido por

$$\langle f, g\rangle = \int_{I^n} f(\mathbf{x})g(\mathbf{x})d\mathbf{x} \text{ para todas } f, g \in \mathcal{C}(I^n)\,,$$

es un hecho conocido que el espacio $(\mathcal{C}(I^n), \langle\cdot,\cdot\rangle)$ es un espacio de Hilbert. Aplicando el teorema de representación de Riesz al funcional L, se tiene que existe $\psi \in \mathcal{C}(I^n)$ tal que

$$L(h) = \int_{I^n} h(\mathbf{x})\psi(\mathbf{x})d\mathbf{x} \text{ para toda } h \in \mathcal{C}(I^n)\,.$$

Dado que ψ es una función continua en I^n, la medida definida por

$$d\mu(\mathbf{x}) = \psi(\mathbf{x})d\mathbf{x}\,,$$

es finita, signada, regular y de Borel en I^n, luego $\mu \in \mathcal{M}(I^n)$.

Entonces, se puede considerar que el funcional L está definido como

$$L(h) = \int_{I^n} h(\mathbf{x})d\mu(\mathbf{x}) \text{ para } h \in \mathcal{C}(I^n)\,, \tag{2.44}$$

luego, se tiene que $\mu \neq 0$ ya que $L \neq 0$.

Claramente, para cada $\mathbf{w} \in \mathbb{R}^n$ y cada $b \in \mathbb{R}$, se tiene que

$$\sigma(\mathbf{w}^T\mathbf{x} + b) \in \overline{D}\,.$$

Entonces, por definición de L, se tiene que

$$\int_{I^n} \sigma(\mathbf{w}^T\mathbf{x} + b)d\mu(\mathbf{x}) = 0,\,,$$

para cada $\mathbf{w} \in \mathbb{R}^n$ y cada $b \in \mathbb{R}$.

Como estamos asumiendo que σ es una función discriminatoria, esta condición implica que $\mu = 0$, lo cual absurdo, pues contradice (2.44).

En consecuencia, D debe de ser un subespacio denso de $\mathcal{C}(I^n)$. □

Lema 2.18. *Toda función medible acotada y sigmoidal es discriminatoria. En particular, toda función sigmoidal continua es discriminatoria.*

Demostración. Sea σ una función sigmoidal medible y acotada, y sea $\mu \in \mathcal{M}(I^n)$ una medida tal que para todo $\mathbf{y} \in \mathbb{R}^n$ y $\theta \in \mathbb{R}$

$$\int_{I^n} \sigma(\mathbf{y}^T\mathbf{x} + \theta) d\mu(\mathbf{x}) = 0\,.$$

Sean además $\varphi \in \mathbb{R}$ y $\{\lambda_k\}_{k=1}^{\infty}$ una sucesión tal que $\lambda_k \xrightarrow{k\to\infty} +\infty$. Se considera, la sucesión de funciones $\{\sigma_{\lambda_k}\}_{k=1}^{\infty}$ definida por

$$\sigma_{\lambda_k}(\mathbf{x}) = \sigma(\lambda_k(\mathbf{y}^T\mathbf{x} + \theta) + \varphi)$$

La sucesión de funciones converge puntualmente hacia la función

$$\gamma(\mathbf{x}) = \begin{cases} 1 & \text{si } \mathbf{y}^T\mathbf{x} + \theta > 0\,, \\ 0 & \text{si } \mathbf{y}^T\mathbf{x} + \theta < 0\,, \\ \sigma(\varphi) & \text{si } \mathbf{y}^T\mathbf{x} + \theta = 0\,, \end{cases}$$

ya que

$$\sigma(\lambda(\mathbf{y}^T\mathbf{x} + \theta) + \varphi) \begin{cases} \to 1 & \text{si } \mathbf{y}^T\mathbf{x} + \theta > 0 \text{ cuando } \lambda \to +\infty, \\ \to 0 & \text{si } \mathbf{y}^T\mathbf{x} + \theta < 0 \text{ cuando } \lambda \to +\infty, \\ = \sigma(\varphi) & \text{si } \mathbf{y}^T\mathbf{x} + \theta = 0 \text{ para todo } \lambda\,. \end{cases}$$

Para cada $\mathbf{y} \in \mathbb{R}^n$ y $\theta \in \mathbb{R}$ se definen los conjuntos

$$\Pi_{\mathbf{y},\theta} = \{\mathbf{x} \in \mathbb{R}^n \colon \mathbf{y}^T\mathbf{x} + \theta = 0\}\,,$$
$$H_{\mathbf{y},\theta} = \{\mathbf{x} \in \mathbb{R}^n \colon \mathbf{y}^T\mathbf{x} + \theta > 0\}\,.$$

Por hipótesis, para todo $k \in \mathbb{N}$ se cumple que

$$\int_{I^n} \sigma_{\lambda_k} d\mu(\mathbf{x}) = \int_{I^n} \sigma(\lambda_k(\mathbf{y}^T\mathbf{x} + \theta) + \varphi) d\mu(\mathbf{x}) = 0\,,$$

entonces, en virtud del Teorema de la Convergencia Dominada (ver [15]), se tiene que

$$0 = \lim_{k\to\infty} \int_{I^n} \sigma_{\lambda_k}(\mathbf{x}) d\mu(\mathbf{x}) = \int_{I^n} \lim_{k\to\infty} \sigma_{\lambda_k}(\mathbf{x}) d\mu(\mathbf{x}) \tag{2.45}$$

$$= \int_{I^n} \gamma(\mathbf{x}) d\mu(\mathbf{x}) = \sigma(\varphi)\mu(\Pi_{\mathbf{y},\theta}) + \mu(H_{\mathbf{y},\theta})\,, \tag{2.46}$$

y como σ es sigmoidal, $\lim_{\varphi\to\infty} \sigma(\varphi) = 1$, por lo que

$$\mu(\Pi_{\mathbf{y},\theta}) + \mu(H_{\mathbf{y},\theta}) = 0\,. \tag{2.47}$$

Para concluir que σ es una función discriminatoria, hay que probar que la medida nula de los conjuntos $\Pi_{\mathbf{y},\theta}$ y $H_{\mathbf{y},\theta}$ para cada $\mathbf{y}$ implica necesariamente que $\mu \equiv 0$. Este hecho es trivial si μ fuese una medida positiva, pero como μ es una medida positiva, hay que trabajar más.

Para ello, se fija $\mathbf{y} \in \mathbb{R}^n$ y se considera el funcional $F\colon \mathcal{H} \longrightarrow \mathbb{R}$ definido por

$$F(h) = \int_{I^n} h(\mathbf{y}^T\mathbf{x}) d\mu(\mathbf{x})\,,$$

donde $\mathcal{H}$ es el conjunto de las funciones medibles y acotadas en I^n.

Como μ es una medida finita en I^n, F es un funcional lineal y acotado en $L^\infty(\mathbb{R})$.

Si se considera $h(u) = \mathbb{1}_{[\theta,+\infty)}(u)$, la función indicatriz del intervalo $[\theta, +\infty)$, teniendo en cuenta (2.47), se tiene que

$$F(h) = \int_{I^n} h(\mathbf{y}^T\mathbf{x}) d\mu(\mathbf{x}) = \mu(\Pi_{\mathbf{y},-\theta}) + \mu(H_{\mathbf{y},-\theta}) = 0\,.$$

Mediante argumentos similares y aplicando la linealidad del funcional, se demuestra que si h es una función simple (i.e. una suma finita de funciones indicatrices de intervalos), entonces $F(h) = 0$. En consecuencia, puesto que las funciones simples son densas en $L^\infty(\mathbb{R})$, necesariamente se tiene que $F \equiv 0$.

En particular, si para cada $\mathbf{y} \in \mathbb{R}^n$ se toman las funciones

$$s(\mathbf{x}) = \sin(\mathbf{y}^T\mathbf{x})\,,$$
$$c(\mathbf{x}) = \cos(\mathbf{y}^T\mathbf{x})\,,$$

como estas son medibles y acotadas, su imagen por el funcional F es

$$F(c + is) = \int_{I^n} \cos(\mathbf{y}^T\mathbf{x}) + i\sin(\mathbf{y}^T\mathbf{x}) d\mu(\mathbf{x}) = \int_{I^n} \exp(i\mathbf{y}^T\mathbf{x}) d\mu(\mathbf{x}) = 0\,.$$

La expresión anterior muestra que la transformada de Fourier de la medida μ es 0, y en consecuencia, $\mu \equiv 0$ (ver [32]), por lo que se concluye que σ es una función discriminatoria.

□

Corolario 2.19 (Teorema de Aproximación Universal.)**.** *Una red neuronal de arquitectura feed-forward con una única capa oculta de un número suficiente neuronas con funciones de activación sigmoidales y continuas, puede aproximar con precisión arbitraria a cualquier función continua definida en un compacto.*

Demostración. Basta combinar el Teorema 2.17 y el Lema 2.18.

Entonces, para cualquier $\varepsilon > 0$, la densidad del conjunto D del Teorema 2.17 permite garantizar que para cualquier función $f \in \mathcal{C}(I^n)$, existe una red neuronal como la descrita en enunciado que es una combinanción lineal finita de funciones sigmoidales de la forma

$$G(\mathbf{x}, \boldsymbol{\theta}) = \sum_{j=1}^{N} \alpha_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j)\,,$$

tal que

$$|G(\mathbf{x}) - f(\mathbf{x})| < \varepsilon \text{ para todo } x \in I^n\,.$$

Lo que demuestra que con una red neuronal, es posible obtener precisión arbitraria para la aproximación de una función de $\mathcal{C}(I^n)$.

□

Capítulo 3

Experimentos Numéricos

Para concluir esta memoria, se van a realizar unos experimentos sencillos para materializar en la práctica los modelos descritos durante el trascurso de este trabajo.

Para la implementación práctica de las redes neuronales, se ha elegido el lenguaje de programación *Python*, desarrollado por Guido van Rossum y lanzado por primera vez en 1991. *Python* dispone de una amplia variedad de bibliotecas, siendo de especial relevancia para este trabajo *TensorFlow*, una biblioteca de código abierto de diferenciación automática diseñada específicamente por *Google* para construir y entrenar modelos de redes neuronales. Para facilitar la creación rápida y sencilla de modelos de redes neuronales, se empleará la API de *Keras*. La combinación de estas tecnologías proporciona un potente *framework* para el desarrollo y aplicación de métodos numéricos basados en modelos de redes neuronales.

3.1. Experimento 1

Con este experimento, se demostrará empíricamente cómo una red neuronal actúa como aproximador universal. Para este ejemplo, la función que se va a aproximar es (1.34), es decir, el precio Black-Scholes de una opción europea sobre una acción que no paga dividendos.

Para el primer experimento, se fijaron los parámetros de la tasa libre de riesgo $r = 0.05$ y la volatilidad $\sigma = 0.1$, y como conjunto de entrenamiento para la red, se utilizó una malla cuadrada equiespaciada de 100×100 nodos en $(S, T) \in [0.8, 1.2] \times [0, 1]$. En esta malla, la primera dimensión corresponde a los valores de las cotizaciones de la acción y la segunda al tiempo hasta el vencimiento. Para la validación de este experimento, se toma otra malla cuadrada equiespaciada

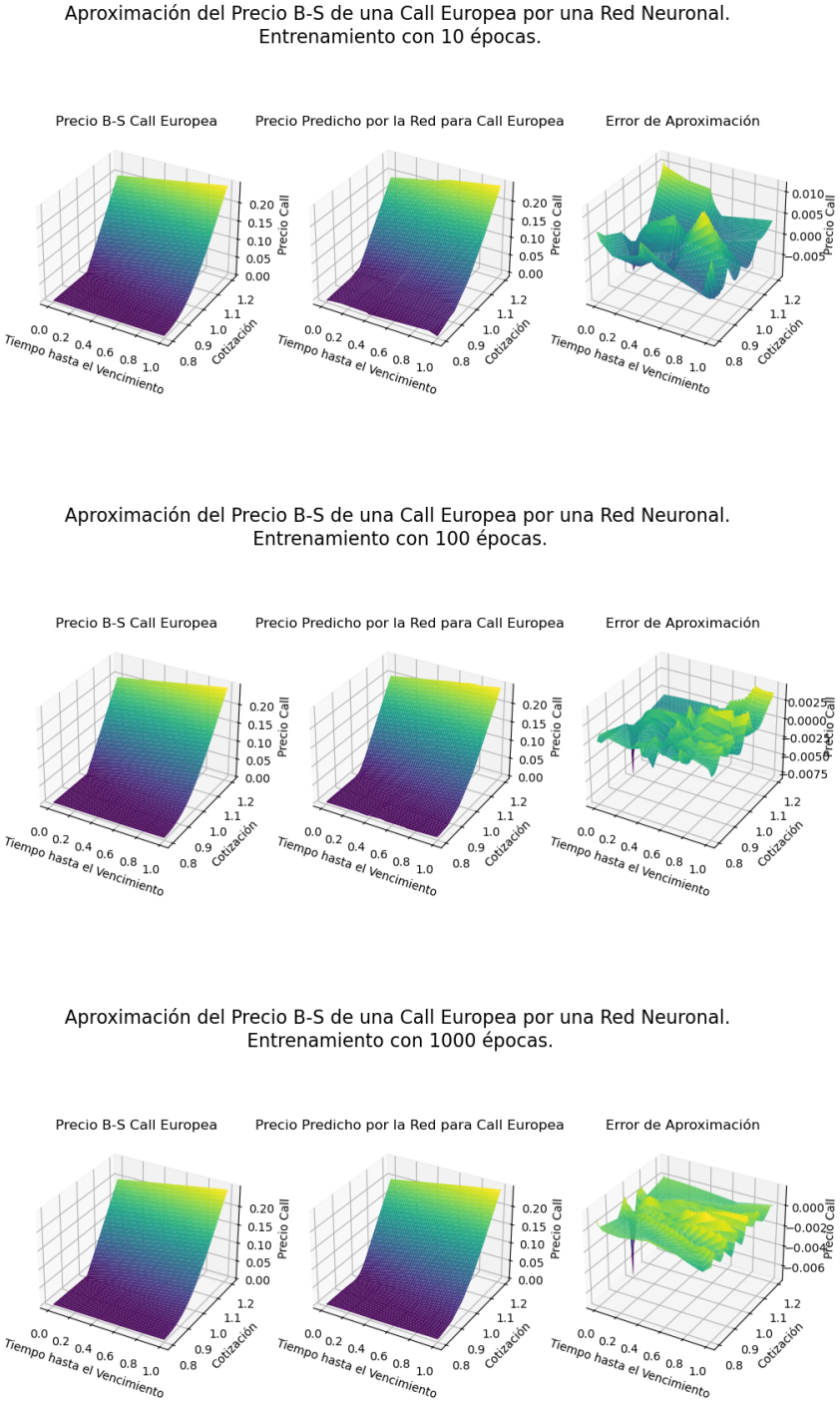


Figura 3.1: Salidas de la red para 10, 100 y 1000 épocas.

de 200×200 nodos en $[0.8, 1.2] \times [0, 1]$ (i.e. el doble de fina que la malla de entrenamiento). Con la red neuronal entrenada, se realizan predicciones sobre esta malla y se comparan con los precios exactos para obtener el error de la aproximación.

En la Figura 3.1, se ha representado el precio exacto de Black-Scholes (izquierda), el precio predicho por la red (centro) y la diferencia entre ambos precios (derecha) para distinto número de épocas (filas). Se observa cómo la calidad de la aproximación mejora con el incremento del número de épocas de test. Con 10 épocas, la red neuronal aproxima la forma de la superficie de precios, pero muestra un error de aproximación considerable. A medida que se incrementa el número de épocas a 100 y 1000, las predicciones de la red se ajustan mejor al precio teórico, reduciendo significativamente el error de aproximación, aunque la ganancia entre 100 y 1000 épocas no es significativa.

Una vez comprobado que la red funciona para dos entradas, se entrena otra red para todos los parámetros del modelo de Black-Scholes. La nueva red tendrá cuatro variables de entrada (el precio de la Call es lineal respecto a la relación $\frac{S}{K}$). Para generar los datos de entrenamiento de esta nueva red red, se ha generado una malla con todas las combinaciones posibles de los siguientes parámetros: $S \in [0.8, 1.15]$, $T \in [0, 1]$, $r \in [0.01, 0.05]$, $\sigma \in [0.05, 0.2]$, tomando 100 valores equiespaciados para S y T y 10 valores equiespaciados para r y σ.

Para cada combinación de los parámetros mencionados, se ha calculado el precio de una Call según el modelo de Black-Scholes utilizando una función de Python que implementa (1.34).

Como conjunto de entrenamiento se selecciona al azar el 70 % de las combinaciones anteriores de los parámetros con los precios y como conjunto de validación se toma el otro 30 % restante, práctica habitual en el entrenamiento de redes (ver [21]).

Con este conjunto de entrenamiento y validación, se ha entrenado durante 1000 épocas un modelo secuencial de Keras formado por una capa de input de tamaño 4, 2 capas ocultas con 20 neuronas cada una y una capa de output de tamaño 1.

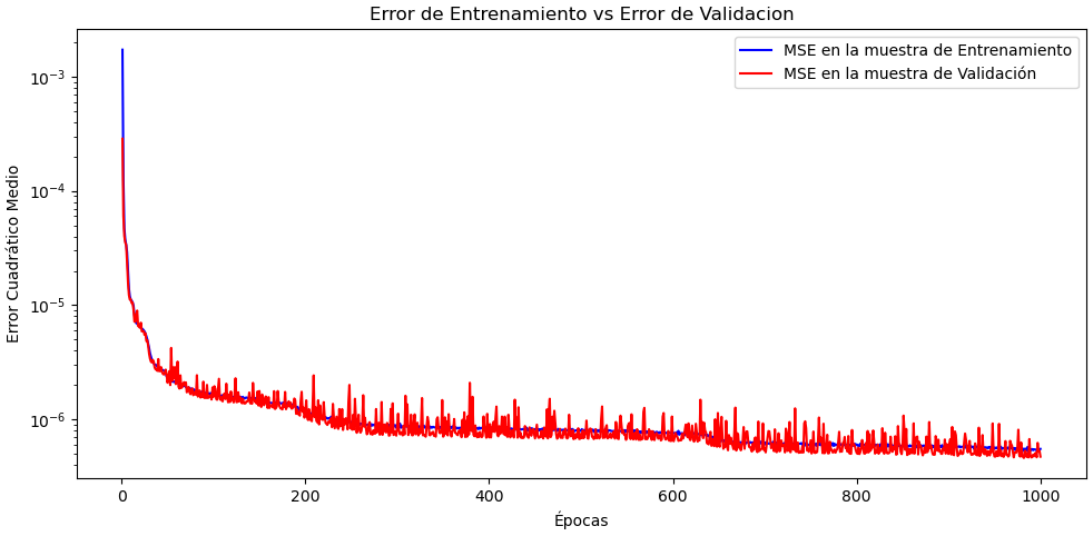


Figura 3.2: Error cuadrático medio de las muestras de entrenamiento y validación en cada época.

La Figura 3.2 muestra cómo se ha desarrollado el entrenamiento de la red, mostrando en escala semi-logarítmica, en azul el error cuadrático medio de la muestra de entrenamiento y en rojo el de la muestra de validación o *test*. Se observa cómo durante las primeras 50 épocas, los errores de ambas muestras tienen comportamientos similares.

Después de este punto, el error en la muestra de validación comienza a fluctuar, mientras que el error en la muestra de entrenamiento continúa disminuyendo, aunque de forma muy tenua. A partir de aproximadamente la época 200, el error de la muestra de entrenamiento se estabiliza, mientras que el de validación muestra mayores fluctuaciones. Por lo visto, entrenar más allá de las 200 épocas parece no ofrecer una mejora significativa.

A continuación, se realizará un análisis del error de aproximación que comete la red.

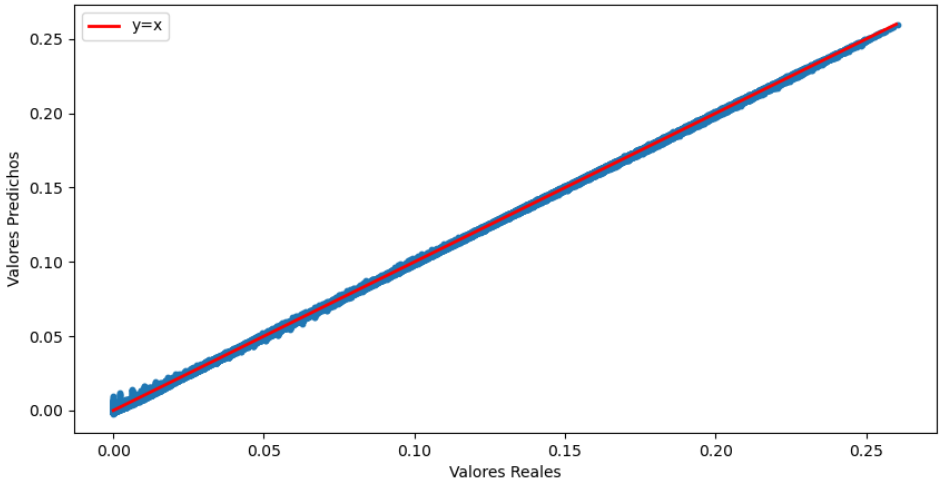


Figura 3.3: Precios reales frente a los precios predichos por la red.

En la Figura 3.3, la coordenada x de cada punto es el precio Black-Scholes de cada elemento del conjunto de test, mientras que la coordenada y es el precio predicho por la red. Los puntos azules representan estas combinaciones de precios reales y predichos. La línea roja, que corresponde a la recta $y = x$, que indicaría una predicción perfecta donde los valores predichos coinciden exactamente con los valores reales.

La proximidad que se observa en la Figura 3.3 de los puntos azules a la recta $y = x$, sugiere una alta precisión de las predicciones de la red, ya que la mayoría de los puntos se alinean estrechamente con la línea de referencia. Esto demuestra que la red neuronal es capaz de aproximar con precisión los precios calculados mediante el modelo de Black-Scholes para la muestra de validación.

La dispersión de algunos puntos alrededor de la línea indica pequeños errores de predicción, que son esperables, pero para garantizar que la aproximación de la red no introduce ningún tipo de sesgo, es interesante estudiar la distribución empírica que tienen los errores de las salidas de la red.

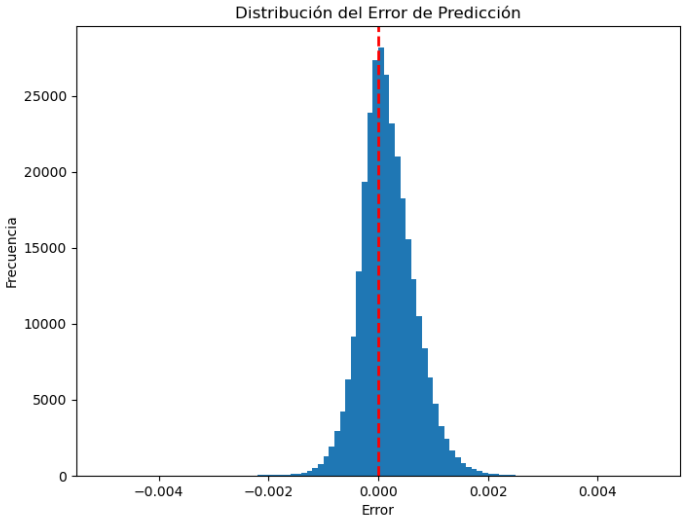


Figura 3.4: Distribución empírica del error de predicción de la red.

En la Figura 3.4 se muestra un histograma con con el error en el eje horizontal y la frecuencia en el eje vertical. Se observa que la distribución del error parece ser simétrica y centrada en cero, lo que sugiere que la red no introduce sesgos de manera sistemática en sus predicciones. La mayoría de los errores son pequeños, lo que indica que la red tiene un buen rendimiento en términos de precisión de predicción.

Por último en la Figura 3.5 se representan los precios de Black-Scholes en función de las cotizaciones y vencimientos. Se ha representado el valor exacto de Black-Scholes (izquierda), el precio predicho por la red (centro) y la diferencia entre ambos (derecha) para los valores de $r = 0.01$ y $\sigma = 0.1$ (primera fila), los valores $r = 0.03$ y $\sigma = 0.15$ (segunda fila) y $r = 0.05$ y $\sigma = 0.2$ (tercera fila). Como se puede apreciar, el error de aproximación es similar en todas las combinaciones de parámetros que se han elegido.

3.2. Experimento 2

En la realidad, ningún modelo teórico (Black-Scholes, Heston, etc.) consigue ajustar perfectamente los precios observados en los mercados. Lo que se intenta es encontrar los modelos (o el valor de los parámetros de los mismos) que mejor ajusten a los precios observados (i.e. calibrar el modelo)

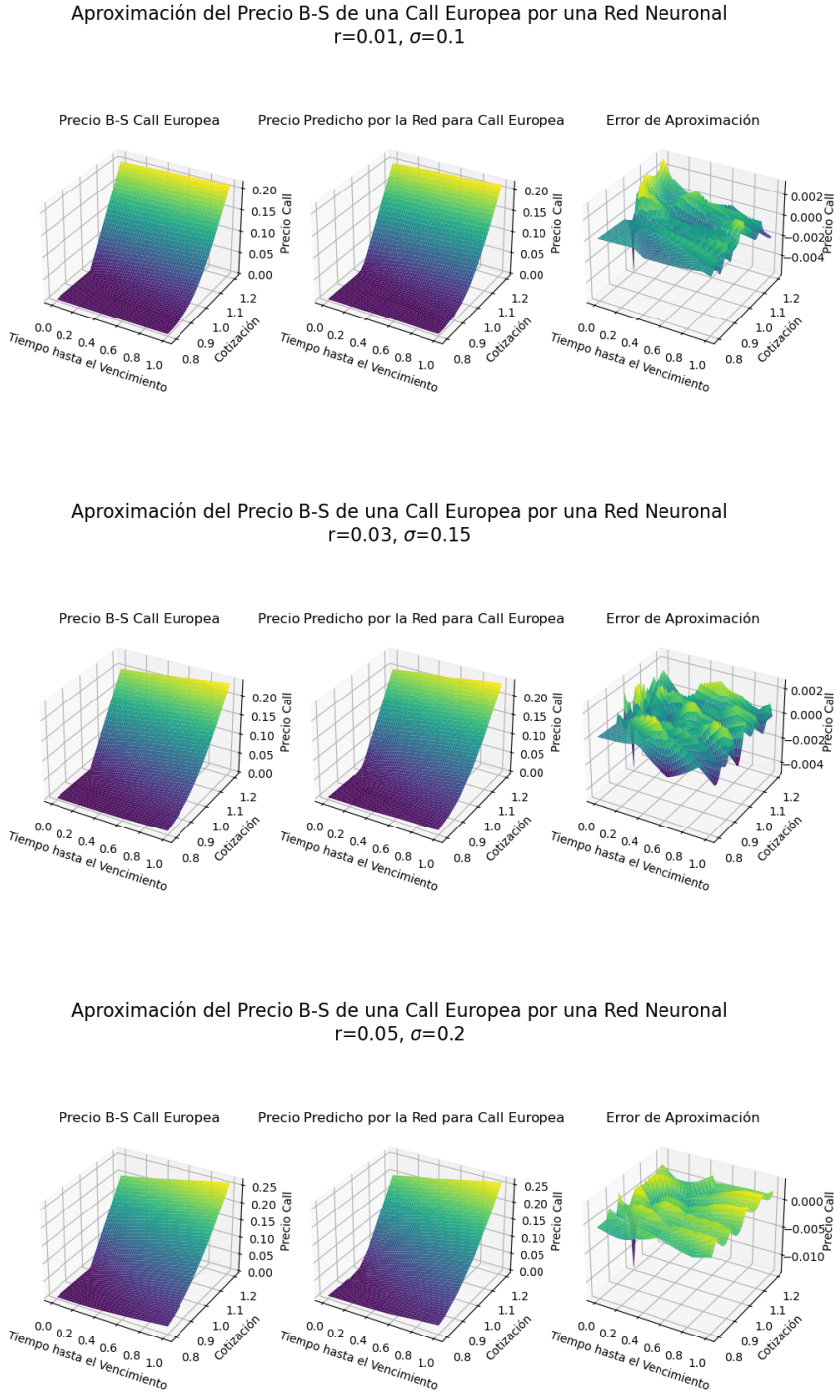


Figura 3.5: Salidas de la red para diferentes combinaciones de parámetros del modelo de Black-Scholes.

En este experimento, se implementará una técnica usada en la valoración de derivados financieros que aparece en [9].

El objetivo es, asumiendo que los precios negociados de los derivados siguen el modelo teórico elegido (o al menos una dinámica muy similar), estimar los parámetros del modelo mediante un ajuste de mínimos cuadrados entre los precios negociados y los precios predichos por el modelo.

En la práctica, se elige un conjunto de datos históricos del pasado reciente y se buscan los parámetros del modelo que mejor ajustan (conocido como el ajuste *in-the-sample*). Posteriormente, se utiliza el valor estimado de esos parámetros para predecir el valor de nuevos contratos (conocido como la predicción *out-the-sample*) tras variaciones, por ejemplo, en la cotización del activo o en las condiciones del contrato como el vencimiento.

Para llevar a cabo este experimento, se utilizará el precio exacto del modelo de Black-Scholes, la segunda red del experimento anterior, que se llamará Modelo 1, y una nueva red, que se llamará Modelo 2. La red del Modelo 2 ha sido entrenada de manera similar a la anterior, con el mismo conjunto de datos, pero con 5 neuronas en cada capa oculta en lugar de 20. El Modelo 2 se ha incorporado para ver si existen diferencias significativas o habría bastado con una red más pequeña.

A falta de datos reales de mercado, se van a realizar dos subexperimentos con datos creados artificialmente. En primer lugar se van a generar los precios de Black-Scholes de 20 contratos fijando $r = 0.03$ y $\sigma = 0.1$ y tomando 5 valores equiespaciados para $S \in [0.85, 1.1]$ y 4 valores equiespaciados para $T \in [0, 1]$ (muestra *in-the-sample* Black-Scholes).

Aunque en el Experimento 1 se comprobó que el error obtenido con la red era pequeño, parece interesante comprobar que la red entrenada es de hecho consistente para estimar los parámetros del mercado en el propio modelo en el que ha sido entrenada.

	Parámetros Estimados	RMSE *in-the-sample*
B-S exacto	$r = 0.030004$, $\sigma = 0.099999$	1.5×10^{-6}
Modelo 1	$r = 0.031929$, $\sigma = 0.096696$	0.000469
Modelo 2	$r = 0.01$, $\sigma = 0.188167$	0.011872

Tabla 3.1: Resultados de la estimación *in-the-sample* con datos Black-Scholes.

En la Tabla 3.1 se puede observar en la primera columna el valor de los parámetros obtenidos con el modelo exacto de Black-Scholes, con la red más grande del Modelo 1 y la red más pequeña del Modelo 2 y en la segunda columna la raíz

del error cuadrático medio (RMSE) cometido con cada modelo en el ajuste *in-the-sample*. Para la muestra *out-the-sample*, se generan otros 20 contratos tomando 5 valores equiespaciados para $S \in [0.83, 1.13]$ y 4 valores equiespaciados para $T \in [0.1, 0.9]$ y los mismos valores $r = 0.03$ y $\sigma = 0.1$. En la Tabla 3.2 se incluyen, para los distintos modelos, el RMSE cometido empleando los parámetros estimados en el análisis *in-the-sample*.

	RMSE *out-the-sample*
B-S exacto	1.49×10^{-6}
Modelo 1	0.000638
Modelo 2	0.011249

Tabla 3.2: Error *out-the-sample* con datos Black-Scholes.

A la vista de los resultados presentados en las Tablas 3.1 y 3.2 se ve que la aproximación de los parámetros mediante el precio exacto de Black-Scholes es bastante buena, al igual que la obtenida por el Modelo 1 que es también aceptable. Los resultados del Modelo 2 difieren mucho más, así que seguramente la red del Modelo 2 es excesivamente pequeña para el objetivo buscado. De hecho, para el Modelo 2, la búsqueda de los parámetros lleva a valores que están fuera del dominio donde la red ha sido entrenada, haciendo que los resultados obtenidos con este modelo no sean fiables.

La segunda parte del experimento va a intentar recrear las condiciones reales de mercado como en [9]. El mercado real no sigue el modelo de Black-Scholes y, debido a la carencia de datos reales de mercado, se va a intentar replicar esas condiciones generando los contratos para el análisis *in-the-sample / out-the-sample* con el modelo de Heston, aunque la estimación de parámetros se va a realizar con el modelo de Black-Scholes.

Para la generación de los datos del modelo de Heston, se han fijado los parámetros $r = 0.02, v = 0.05, \theta = 0.03, \sigma = 1.3, \rho = -0.7$ y $(\kappa + \lambda) = 1$ y se han tomado las mismas cotizaciones y vencimientos a los tomados en el experimento anterior.

	Parámetros Estimados	RMSE *in-the-sample*
B-S exacto	$r = 0.045899$, $\sigma = 0.166185$	0.002816
Modelo 1	$r = 0.047801$, $\sigma = 0.164300$	0.003031
Modelo 2	$r = 0.05$, $\sigma = 0.124869$	0.009568

Tabla 3.3: Resultados de la estimación *in-the-sample* con datos Heston.

	RMSE *out-the-sample*
B-S exacto	0.002768
Modelo 1	0.003161
Modelo 2	0.009655

Tabla 3.4: Error *out-the-sample* con datos Heston.

En las Tablas 3.3 y 3.4, se aprecia como el modelo exacto de Black-Scholes comete un error mucho mayor que en las Tablas 3.1 y 3.2, cosa esperable ya que se está aproximando un modelo por otro mucho más simple. Los resultados obtenidos con la red del Modelo 1 de nuevo son bastante consistentes y más similares a los resultados de Black-Scholes en comparación con la red del Modelo 2, que de nuevo tiende a salirse del dominio en el que ha sido entrenada.

3.3. Conclusiones

Este TFG ha sido un trabajo exploratorio en dos sentidos. En primer lugar ha sido una toma de contacto con el campo de las Matemáticas Financieras, los modelos que se emplean y las propiedades que intentan recrear a nivel teórico (así como las carencias que presentan). Por el otro, ha sido una toma de contacto con las Redes Neuronales y sus potencialidades.

En los experimentos realizados se ha visto cómo una red neuronal puede aproximar bastante bien un modelo de valoración, y que sólo han hecho falta gran cantidad de datos de entrenamiento previamente calculados. Así mismo, la evaluación de una red neuronal entrenada en modelos más complejos, es mucho menos demandante computacionalmente que el cálculo de un precio a través de un método numérico clásico.

Esta eficiencia es la gran potencialidad de las redes en este campo, ya que permite la valoración en tiempo real de derivados financieros, una capacidad que es crucial para la toma de decisiones rápidas y precisas en los mercados financieros.

Existe una amplia literatura del empleo de las Redes Neuronales en este campo. Por ejemplo, se pueden emplear redes neuronales entrenadas con datos de mercado [3] o, como se menciona en [2], las redes neuronales pueden integrarse con algoritmos tradicionales para acelerar significativamente el proceso de cálculo de los precios de los derivados, proporcionando resultados más rápidos y eficientes.

Otra forma de aplicar las redes neuronales en el contexto de las Finanzas Cuantitativas es mediante el uso de un tipo de redes conocidas como Physics-Informed Neural Networks (PINN's). Estas redes están diseñadas para aproximar

procesos que pueden describirse mediante en derivadas parciales, lo que las hace particularmente útiles para aproximar modelos multidimensionales en finanzas que siguen tales ecuaciones.

En resumen, los temas abordados en esta sección sugieren áreas de estudio futuras muy interesantes y, el desarrollo de este trabajo ha sido una experiencia extremadamente enriquecedora, que me ha proporcionado una cierta comprensión y ganas de profundizar sobre la modelización de los mercados y la aplicación de las redes neuronales en la valoración de derivados financieros.

Bibliografía

[1] Achdou, Y., and Pironneau, O. Finite element methods for option pricing. *Université Pierre et Marie Curie* (2007), 1–12.

[2] Anderson, D., and Ulrych, U. Accelerated american option pricing with deep neural networks. *Quantitative Finance and Economics 7*, 2 (2023), 207–228.

[3] Andreou, P. C., Charalambous, C., and Martzoukos, S. H. Pricing and trading european options by combining artificial neural networks and parametric models with implied parameters. *European Journal of Operational Research 185*, 3 (2008), 1415–1433.

[4] Baladram, M. S., Koike, A., and Yamada, K. Introduction to supervised machine learning for data science. *Interdisciplinary information sciences 26*, 1 (2020), 87–121.

[5] Billingsley, P. *Probability and measure*. John Wiley & Sons, 2017.

[6] Björk, T. *Arbitrage theory in continuous time*. Oxford university press, 2009.

[7] Black, F., and Scholes, M. The pricing of options and corporate liabilities. *Journal of political economy 81*, 3 (1973), 637–654.

[8] Capiński, M., Kopp, E., and Traple, J. *Stochastic Calculus for Finance*. Mastering Mathematical Finance. Cambridge University Press, 2012.

[9] Christoffersen, P., and Jacobs, K. Which garch model for option valuation? *Management science 50*, 9 (2004), 1204–1221.

[10] Cohen, S. N., and Elliott, R. J. *Stochastic calculus and applications*, vol. 2. Springer, 2015.

[11] Cox, J. C., Ingersoll Jr, J. E., and Ross, S. A. A theory of the term structure of interest rates. *Econometrica 53*, 2 (1985), 385–407.

[12] Cross, S. S., Harrison, R. F., and Kennedy, R. L. Introduction to neural networks. *The Lancet 346*, 8982 (1995), 1075–1079.

[13] Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems 2*, 4 (1989), 303–314.

[14] Elliott, R. J., and Kopp, P. E. *Mathematics of financial markets*, vol. 10. Springer Science & Business Media, 2005.

[15] Galindo Soto, F., Sanz Gil, J., and Tristán Vega, L. A. *Guía práctica de cálculo infinitesimal en varias variables*. Ediciones Paraninfo, 2005.

[16] Gatheral, J. *The volatility surface: a practitioner's guide*. John Wiley & Sons, 2011.

[17] Gil-Pelaez, J. Note on the inversion theorem. *Biometrika 38*, 3-4 (1951), 481–482.

[18] Glasserman, P. *Monte Carlo methods in financial engineering*, vol. 53. Springer, 2004.

[19] Heston, S. L. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies 6*, 2 (1993), 327–343.

[20] Higham, C. F., and Higham, D. J. Deep learning: An introduction for applied mathematicians. *Siam review 61*, 4 (2019), 860–891.

[21] Hirsa, A., Karatas, T., and Oskoui, A. Supervised deep neural networks (dnns) for pricing/calibration of vanilla/exotic options under various different processes, 2019.

[22] Jarrow, R., and Protter, P. A short history of stochastic integration and mathematical finance: the early years, 1880-1970. *Lecture Notes-Monograph Series* (2004), 75–91.

[23] Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks 6*, 6 (1993), 861–867.

[24] Linetsky, V. Spectral methods in derivatives pricing. *Handbooks in Operations Research and Management Science 15* (2007), 223–299.

[25] Longstaff, F. A., and Schwartz, E. S. Valuing american options by simulation: a simple least-squares approach. *The review of financial studies 14*, 1 (2001), 113–147.

[26] Merton, R. C. Theory of rational option pricing. *The Bell Journal of economics and management science* (1973), 141–183.

[27] Montavon, G. Introduction to neural networks. *Machine learning meets quantum physics* (2020), 37–62.

[28] Ndiaye, M. The riccati equation, differential transform, rational solutions and applications. *Applied Mathematics 13*, 9 (2022), 774–792.

[29] Nguyen, L. M., Nguyen, N. H., Phan, D. T., Kalagnanam, J. R., and Scheinberg, K. When does stochastic gradient algorithm work well? *arXiv preprint arXiv:1801.06159* (2018).

[30] Nielsen, M. A. Neural networks and deep learning, 2018.

[31] Oksendal, B. *Stochastic differential equations: an introduction with applications*. Springer Science & Business Media, 2013.

[32] Rudin, W. *Functional Analysis*. McGraw-Hill, New York, 1973.

[33] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature 323*, 6088 (1986), 533–536.

[34] Sharma, S., Sharma, S., and Athaiya, A. Activation functions in neural networks. *Towards Data Sci 6*, 12 (2017), 310–316.

[35] Simon, H. *Neural networks and learning machines*. Pearson Education, Inc, 2009.

[36] Soner, H. M., Shreve, S. E., and Cvitanic, J. There is no nontrivial hedging portfolio for option pricing with transaction costs. *The Annals of Applied Probability 5*, 2 (1995), 327–355.

[37] Wilmott, P. *Paul Wilmott on Quantitative Finance*. Wiley, 2010.