Universidad de Guanajuato

Análisis Basado en Matrices Aleatorias y Probabilidad Libre de la Dinámica de Generalización de Redes Neuronales

T E S I S

Para obtener el título de

Licenciado en Matemáticas

P R E S E N T A:

Carlos Misael Madrid Padilla

Directores de Tesis:

Dr. Mario Alberto Diaz Torres

Dr. Víctor Manuel Pérez Abreu Carrión

GUANAJUATO, GTO Junio, 2019

Agradecimientos

A mi familia por todo el apoyo que me han brindado a lo largo de mi vida. Especialmente a mi mamá y mi papá que, aunque ya no está entre nosotros, siempre formará parte de mis logros, a mis hermanos Oscar y Jose. A Lizda por su incondicional amor y a mi hija Erandi que es el motor de mi vida. Agradezco a mi suegra Lourdes por aceptarme en su familia y ser mi segunda madre.

Quiero agradecerle a mis directores de tesis, Dr Víctor Pérez Abreu y Dr Mario Diaz. Muchas gracias por su paciencia, sus valiosos consejos académicos y personales, y sobre todo muchas gracias por su amistad. Gracias por ser un ejemplo a seguir y guiarme en mi formación académica.

Doy gracias a mis sinodales el Dr. Rogelio Ramos Quiroga, Dr. Carlos Vargas Obieta y M.C. Ricardo José Guerrero Rodríguez por sus correcciones y sugerencias para lograr que la tesis saliera lo mejor posible.

Índice general

Introducción

Las redes neuronales artificiales son un campo muy importante dentro del aprendizaje automático. Se basan en la idea de combinar ciertos parámetros para predecir resultados. Estos parámetros se buscan de forma algorítmica y comprender la dinámica de la combinación de estos parámetros, proceso conocido como entrenamiento de la red neuronal artificial, es uno de los temas clave para la mejora de algoritmos de optimización, así como para la comprensión teórica de porqué las redes neuronales artificiales funcionan tan bien hoy en día.

Un grupo de investigadores del área de computación en la universidad de Stanford, Saxe et al. [17], en 2013 mostraron empíricamente que las redes neuronales artificiales usadas en la práctica exhiben el mismo comportamiento en su dinámica de la combinación de los parámetros, que un tipo de redes neuronales más simples, las llamadas redes neuronales artificiales lineales. En 2018, los ingenieros en el departamento de estadística y señales de la universidad Paris-Saclay, Liao y Couillet [18] consideran una red neuronal artificial lineal con el objeto de realizar un análisis de la dinámica de entrenamiento y generalización. Su trabajo hace uso de herramientas de la teoría de matrices aleatorias.

El objetivo de esta tesis es presentar una exposición detallada del trabajo de [18]. Nuestra exposición da pruebas rigurosas de los principales resultados en [18], así como pruebas diferentes usando la teoría de probabilidad libre, en particular el concepto de distribución infinitesimal y su libertad, lo cual da claridad a ciertos aspectos mostrados en [18].

La estructura de la tesis es la siguiente. El Capítulo 1 está dedicado a los conceptos y resultados de matrices aleatorias y probabilidad libre que se usan en el resto de esta tesis. Comenzamos con las definiciones básicas de teoría de matrices aleatorias y posteriormente con el enfoque de probabilidad libre a la teoría de matrices aleatorias. Para finalizar, se da la definición de distribución infinitesimal, la cual es una extensión de probabilidad libre.

En el Capítulo 2 se presenta el modelo de red neuronal artificial lineal de una capa. Describiremos a detalle el análisis que se realiza en [18] sobre el entrenamiento de la red neuronal artificial. Específicamente, se realiza una aproximación a tiempo continuo del algoritmo descenso de gradiente, en donde surge de forma natural la matriz de covarianza muestral de los datos de entrenamiento. De esta forma, con lo que se presenta en el Capítulo 1, se trata de entender el qué tan bien la red neuronal trabaja con datos que no son de entrenamiento, lo cual es conocido como generalización de la red neuronal.

El Capítulo 3 se enfoca en la demostración a detalle del resultado central del trabajo

de Liao y Couillet [18]. En la primera sección establecemos una serie de lemas técnicos que formalizan algunas observaciones en [18], y damos una rigurosa explicación del uso de equivalentes deterministicos para obtener el comportamiento asintótico de la generalización de la red neuronal (Teorema 3.1.1). En la segunda parte, proporcionamos una prueba rigurosa de este resultado que además de equivalentes deterministicos utiliza distribución infinitesimal. Finalmente, en la última sección se da la implementación computacional que ilustra el resultado principal en esta tesis.

Con el objetivo de hacer el presente trabajo autocontenido, al final del mismo se incluyen tres apéndices. El Apéndice A contiene los resultados principales sobre probabilidad clásica. Mientras que el Apéndice B proporciona las identidades de matrices y funciones que se utilizan en esta tesis. Finalmente el Apéndice C contiene los códigos computacionales de la realización de la red neuronal artificial.

Capítulo 1

Preliminares Sobre Matrices Aleatorias y Probabilidad Libre

En este capítulo se presentan elementos básicos que se utilizarán en la tesis más adelante, así como nociones generales de matrices aleatorias. Primero se dan las definiciones y resultados de teoría de matrices aleatorias, en los cuales se centra el análisis realizado en este trabajo. Posteriormente, se da brevemente teoría sobre probabilidad libre, enfocada específicamente en matrices aleatorias. Por último se presenta una extensión de la teoría de probabilidad libre, conocida como distribución infinitesimal, con la cual al final del Capítulo 3 se dará un enfoque distinto a la demostración del resultado principal en [18].

1.1. Teoría de Matrices Aleatorias

Las matrices aleatorias son hoy en día una de las áreas de investigación más vigorosa, activas y relevantes. Además tiene considerables aplicaciones en problemas retadores de vanguardia dentro de otras disciplinas como machine learning (redes neuronales artificiales), sistemas de comunicación inalámbrica, estadística, entre otras áreas.

La Teoría de Matrices Aleatorias (Random Matrix Theory) fue introducida por primera vez en la estadística matemática por Wishart [1] en 1928. El tema ganó prominencia cuando Wigner [2] introdujo el concepto de distribución estadística de los niveles de energía nuclear en la década de los 50s. A manera de difusión del tema presentamos elementos básicos de esta teoría, más allá de lo usado en está tesis.

El siguiente material se puede encontrar en [3] y [4]. Otra referencia es [5].

1.1.1. Definiciones Básicas

Denotamos por $\mathbb{M}_{p\times n}(\mathbb{F})$ el espacio lineal de matrices $p \times n$ en el campo $\mathbb{F}$. Se estará considerando $\mathbb{F} = \mathbb{R}$ ó $\mathbb{F} = \mathbb{C}$, además de simplificar la escritura de $\mathbb{M}_{p\times n}(\mathbb{F})$ por $\mathbb{M}_p(\mathbb{F})$ cuando $p = n$. Se usará $\mathcal{B}(\mathbb{M}_{p\times n}(\mathbb{F})) = \mathcal{B}(\mathbb{M}_{p\times n})$ para los borelianos.

Observación. Se puede identificar a $\mathbb{M}_{p\times n}(\mathbb{C})$ como $\mathbb{R}^{2(p\times n)}$ y $\mathbb{M}_{p\times n}(\mathbb{R})$ como $\mathbb{R}^{p\times n}$.

Definición 1.1.1. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad y $\mathcal{Q} \subset \mathbb{M}_{p\times n}(\mathbb{F})$, $\mathcal{Q} \in \mathcal{B}(\mathbb{M}_{p\times n})$. Una función $X : \Omega \longrightarrow \mathcal{Q}$ es una matriz aleatoria si

$$X^{-1}(A) \in \mathcal{F}, \quad \forall A \in \mathcal{B}(\mathcal{Q}) = \mathcal{Q} \cap \mathcal{B}(\mathbb{M}_{p\times n}).$$

La distribución en $\mathcal{Q}$ de X es la medida de probabilidad

$$\mu_X(A) = \mathbb{P}(X \in A), \quad A \in \mathcal{B}(\mathcal{Q}).$$

De igual manera que en el caso de vectores aleatorios; por un resultado de teoría de la medida, si $X = (X_{ij}) \in \mathcal{Q}$, entonces X es matriz aleatoria con valores en $\mathcal{Q}$ si, y solo si, X_{ij} es una variable aleatoria (compleja o real) para todo $i = 1, ...p$ y $j = 1, ..., n$.

Definición 1.1.2. Decimos que la matriz aleatoria X en Q tiene densidad, si existe una función $f : Q \to [0, \infty)$ tal que

$$\mathbb{P}(X \in A) = \int_{Q\cap A} f(x)dx, \quad \forall A \in \mathcal{B}(Q)$$

donde $dx = \prod_{i=1}^{m}$ con $Q \cong \mathbb{R}^m$ y además

$$\int_Q f(x)dx = 1.$$

Ejemplo 1.1.3. Sea $Z = (Z_{ij}) \in \mathbb{M}_{p\times n}(\mathbb{R})$, con $Z_{ij} \sim \mathcal{N}(0, 1)$ variables aleatorias independientes para $i = 1, ..., p$, $j = 1, ..., n$. Entonces, Z tiene densidad

$$f_Z(z) = \left(\frac{1}{\sqrt{2\pi}}\right)^{pn} e^{\frac{-1}{2}||z||^2}, \quad \forall z \in \mathbb{M}_{p\times n}(\mathbb{R}).$$

En el presente trabajo son de interés las matrices $\mathcal{Q} = \mathbb{H}_p(\mathbb{C}) \cong \mathbb{R}^{p^2}$ y $\mathcal{Q} = \mathbb{S}_p(\mathbb{R}) \cong \mathbb{R}^{\frac{p(p+1)}{2}}$, es decir, las matrices hermitianas y matrices simétricas respectivamente.

Definición 1.1.4. Si X es una matriz aleatoria en $\mathbb{M}_p(\mathbb{C})$ con entradas independientes, la llamamos matriz de Ginibre.

Las siguientes matrices son de las matrices mas famosas en la teoría de matrices aleatorias y aparecen en la mayoría de las aplicaciones de esta teoría.

Definición 1.1.5. Se dice que $X = (X_{ij}) \in \mathbb{M}_p(\mathbb{C})$ es matriz de Wigner si

i) (Caso real) X es real simétrica y $\{X_{ij} : 1 \leq i \leq j \leq p\}$ son variables aleatorias independientes.

ii)(Caso complejo) X es hermitiana compleja y $\{X_{ij} : 1 \leq i \leq j \leq p\}$ son variables aleatorias independientes.

En ocasiones, se pide para una matriz de Wigner que tenga segundo momento finito y que encima de la diagonal tenga un medio de la varianza de la diagonal. También es común pedir segundo momento y simetría de las distribuciones.

Las matrices de Wigner juegan un papel importante dentro de la física nuclear y la física matemática. Mencionamos también que las matrices de Wigner tienen un significado estadístico fuerte, pues bajo ciertas condiciones satisfacen la ley del semicírculo.

Definición 1.1.6. Sea $X_1, ..., X_n$ vectores independientes con distribución $\mathcal{N}(0_p, \Sigma_p)$. Consideremos la matriz de tamaño $p \times n$, $X = [X_1, .., X_n]$. Decimos que la matriz $M = XX^T$ tiene distribución Wishart, mientras que M es una matriz de Wishart con n grados de libertad y matriz de covarianza Σ. M es denotada por $W_p(n, \Sigma)$.

Cuando $X_i \sim \mathcal{N}(0_p, I_p)$, M es llamada la matriz Wishart de n grados de libertad. Como ya se había mencionado esta es la primera matriz aleatoria en la historia, 1928. El caso 2×2 lo estudió Fisher.

Más adelante, en el Capítulo 2, veremos que una normalización de una traslación de la matriz Wishart surge naturalmente en el análisis realizado en esta tesis.

1.1.2. Ensambles de Matrices

Aparecieron por primera vez en [6], introducidos por Wigner.

Definición 1.1.7. Un ensamble $(X_m)_{m\geq 1}$ es una sucesión de matrices aleatorias tal que para todo $m \geq 1, X_m$ es una matriz $m \times m$.

Los ensambles más estudiados son los gaussianos:
Sea Z matriz $p \times p$ Ginibri gaussiana en $\mathbb{M}_p(\mathbb{R})$. Definimos $G \in \mathbb{S}_p$ como

$$G = \frac{1}{\sqrt{2}} \left(Z + Z^T \right)$$

si $Z \in \mathbb{M}_p(\mathbb{C})$, definimos $G \in \mathbb{H}_p(\mathbb{C})$ como

$$\hat{G} = \frac{1}{\sqrt{2}} \left(Z + Z^* \right).$$

Definición 1.1.8. (Ensambles especiales)
1) Un ensamble $G = (G^p)_{p\geq 1}$ se dice Gaussiano Ortogonal y se abrevia GOE; si $\forall p \geq 1, G^p = (G^p_{ij})_{i,j=1,...,p}$ es tal que $\{G^p_{ij} : 1 \leq i \leq j \leq p\}$ son variables aleatorias independientes, $G^p \in \mathbb{S}_p(\mathbb{R})$ y

$$G^p_{ii} \sim \mathcal{N}(0, 2),$$
$$G^p_{ij} \sim \mathcal{N}(0, 1) \ \ \forall i \neq j.$$

Es decir G^p es como G.
2) Un ensamble $G = (G^p)_{p\geq 1}$ se dice Gaussiano Unitario y se abrevia GUE; si $\forall p \geq 1, G^p =$

$(G^p_{ij})_{i,j=1,...,p}$ es tal que $\{G^p_{ij} : 1 \leq i \leq j \leq p\}$ son variables aleatorias independientes, $G^p \in \mathbb{H}_p(\mathbb{C})$ y

$$G^p_{ij} \sim \mathcal{N}(0,1).$$

Es decir G^p es como $\hat{G}$.

Las matrices de los ensambles GUE y GOE son matrices de Wigner; lo interesante es que también tenemos un recíproco, que se presenta a continuación.

Teorema 1.1.1. *Sea $X \in \mathbb{S}_p(\mathbb{R})$ matriz aleatoria de Wigner, no diagonal e invariante bajo conjugaciones ortogonales. Entonces X es GOE(p).*

1.1.3. Teorema de Marchenko-Pastur

En la teoría matemática de matrices aleatorias, la distribución de Marchenko-Pastur, o ley de Marchenko-Pastur, describe el comportamiento asintótico de valores singulares de matrices aleatorias rectangulares grandes. El teorema lleva el nombre de los matemáticos ucranianos Vladimir Marchenko y Leonid Pastur que demostraron este resultado en 1967.

Definición 1.1.9. Sea $A \in \mathbb{H}_p(\mathbb{C})$. Entonces los valores propios de A son reales, $\lambda_1, ..., \lambda_p$. Definimos la función de distribución empírica espectral (ESD) de la matriz A como

$$F^A(x) = \frac{1}{p}\#\{j \leq p : \lambda_j \leq x\}.$$

Observemos que no se ha pedido necesariamente que A sea aleatoria. Esto da un modelo distinto al de Kolmogorov, pues a matrices no aleatorias se les da una distribución, es decir, que matrices con los mismos eigenvalores tendrán la misma distribución. La importancia de la ESD es que muchas estadísticas de interés son funcionales de la ESD.

Uno de los principales problemas en Teoría de Matrices Aleatorias es investigar la convergencia de la sucesión de ESDs, $\{F^{A_n}\}$, para una sucesión de matrices aleatorias $\{A_n\}$. La distribución limite, que usualmente es no-aleatoria, es llamada distribución limite espectral (LSD) para la sucesión A_n.

Proposición 1.1.1. *El k-ésimo momento de F^A puede escribirse como,*

$$\beta_{p,k}(A) = \int_{-\infty}^{\infty} x^k F^A(dx) = \frac{1}{p} Tr(A^k).$$

Esta expresión juega un papel fundamental en teoría de matrices aleatorias. Por el teorema de convergencia de momentos, el problema de demostrar que las ESDs de la sucesión A_n converge a otra distribución se reduce a demostrar que, para cada k, la sucesión $\frac{1}{p}tr(A^k)$ tiende a el límite β_k (momento k-ésimo de la LSD).

La transformada de Stieltjes es otra herramienta importante en Teoría de Matrices Aleatorias.

Definición 1.1.10. Sea G una función de distribución. Definimos la transformada de Stieltjes de G como,

$$m_G(z) = \int \frac{1}{x-z} dG(x), \;\; z \in \mathcal{H}^+$$

donde $\mathcal{H}^+ = \{z \in \mathbb{C} : \text{Im} z > 0\}$.

Proposición 1.1.2. *Se tiene que,* $m_{F^A}(z) = \frac{1}{p} Tr(A - zI).^{-1}$

Definición 1.1.11. Sea $A \in \mathbb{H}_p(\mathbb{C})$. El resolvente de la matriz A, denotado por $Q_A(z)$, es la expresión

$$Q_A(z) = (A - zI)^{-1}.$$

El resolvente de una matriz es una técnica para aplicar conceptos del análisis complejo al estudio del espectro de operadores en espacios de Banach y espacios más generales. La justificación formal de las manipulaciones se puede encontrar en el marco del cálculo funcional holomórfico. Una característica interesante es la siguiente,

Proposición 1.1.3. *Sea* $||\cdot||$ *la norma de operador para matrices. Entonces, si* $z = a + ib$ *se tiene que* $||Q_A(z)|| \leq \frac{1}{b}$.

Propiedades concernientes al resolvente de una matriz se encuentran en [9].

Supongamos que $\{x_{jk} : j, k = 1, 2, 3...\}$ es un doble arreglo de v.a.i.i.d complejas con media cero y varianza σ^2. Sea $X_j = [x_{1j}, ..., x_{pj}]^T$ y $X = [X_1, ..., X_n]$.

Definición 1.1.12. La matriz de covarianza muestral esta dada por

$$S_n = \frac{1}{n} X X^*.$$

Observación. Cuando x_{11} tiene distribución $\mathcal{N}(0, 1)$, la matriz nS_n es simplemente la matriz de Wishart con n grados de libertad. En tal caso, se denota a X como la matriz Z con entradas z_{ij}.

La matriz de covarianza muestral es la matriz aleatoria más importante en la inferencia estadística multivariada. Es fundamental en la prueba de hipótesis, análisis de componentes principales, análisis factorial y análisis de discriminación. Muchas estadísticas de prueba están definidas por sus valores propios.

En adelante se estará considerando el caso en que la matriz de covarianza muestral es en realidad una normalización de la matriz de Wishar con n grados de libertad.

Teorema 1.1.2. (Marchenko-Pastur)
Supongamos que $\frac{p}{n} \to c \in (0, \infty)$ *cuando* $n, p \to \infty$. *Entonces con probabilidad 1,* $F^{S_n} \to F^c$ *donde* F^c *es la distribución de Marchenko-Patur dada por*

$$F^c(dx) = (1 - c^{-1})^+ \delta_0(dx) + \frac{1}{2\pi c x} \sqrt{(x-a)^+ (b-x)^+}$$

donde $a = (1 - \sqrt{c})^2, b = (1 + \sqrt{c})^2$ *y* $\delta_0 = 1_0(x)$.

Proposición 1.1.4. *La transformada de Stieltjes de la distribución Marchenko-Pastur esta dada por*

$$m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{(1-c-z)^2-4cz}}{2cz}.$$

Proposición 1.1.5. *La transformada de Stieltjes de la distribución Marchenko-Pastur satisface la ecuación*

$$(zm(z)+1)(cm(z)+1) = m(z).$$

Proposición 1.1.6. *Supongamos que* $\frac{p}{n} \to c \in (0,\infty)$ *cuando* $n, p \to \infty$ *y consideremos* $p = p(n)$. *Entonces cuando* $n \to \infty$,

$$m_{F^{S_n}}(z) - m(z) \to 0$$

con probabilidad 1.

Observación. Esto es, se tiene que $\frac{1}{p}Tr\left(Q_{S_n}(z)\right) - m(z) = \frac{1}{p}Tr\left(S_n - zI\right)^{-1} - m(z)$ converge a 0 con probabilidad 1 cuando $n \to \infty$. Por lo que, el resolvente de la matriz de covarianza muestral esta estrechamente relacionado con la distribución Marchenko-Pastur.

Proposición 1.1.7. *Supongamos que* $\frac{p}{n} \to c \in (0,\infty)$ *cuando* $n, p \to \infty$. *Entonces cuando* $n \to \infty$

$$Q_{S_n}(z)_{1,1} \to \frac{1}{1-c-z-zcm(z)}$$

con probabilidad 1, donde $Q_{S_n}(z)_{1,1}$ *representa la entrada (1,1) de la matriz* $Q_{S_n}(z)$.

Por la Proposición 1.1.5, la Proposición 1.1.7 implica que

$$Q_{S_n}(z)_{1,1} \to m(z)$$

cuando $n \to \infty$, con probabilidad 1.

Antes de finalizar esta subsección se presenta una proposición la cual permite establecer una relación entre el resolvente de la matriz de covarianza muestral y el co-resolvente de esta matriz.

Definición 1.1.13. Sea $AA^* \in \mathbb{H}_p(\mathbb{C})$. El co-resolvente de la matriz AA^*, denotado por $\tilde{Q}_{AA^*}(z)$, es el resolvente de la matriz A^*A.

Proposición 1.1.8. *Sea* $A \in \mathbb{M}_{p\times n}(\mathbb{C})$. $B \in \mathbb{M}_{n\times p}(\mathbb{C})$ *tal que* AB *es hermitiana. Entonces,*

$$m_{F^{BA}}(z) = \frac{p}{n}m_{F^{AB}}(z) + \frac{p-n}{n}\frac{1}{z}.$$

Observación. La Proposición 1.1.8 nos proporciona la relación

$$m_{F^{\frac{1}{n}Z^TZ}}(z) = \frac{p}{n}m_{F^{S_n}}(z) + \frac{p-n}{n}\frac{1}{z}$$

lo cual implica que $m_{F^{\frac{1}{n}Z^TZ}}(z) = \frac{1}{n}Tr(Q_{\frac{1}{n}Z^TZ}(z))$ converge a $\tilde{m}(z)$ con probabilidad 1, cuando $n \to \infty$. Mas aun $\tilde{m}(z)$ satisface la siguiente ecuación

$$\tilde{m}(z) = cm(z) + (c-1)\frac{1}{z}$$

con los mismos supuestos que el Teorema 1.1.2.

En términos del co-resolvente, la Proposición 1.1.8 dice que la traza normalizada del co-resolvente de la matriz de covarianza muestral converge a $\tilde{m}(z)$, con probabilidad 1.

1.1.4. Equivalentes Deterministicos

Las primeras aplicaciones de la teoría de matrices aleatorias en el campo de las comunicaciones inalámbricas, trataron originalmente el comportamiento límite de algunos modelos matriciales aleatorios simples. En particular, estos resultados son atractivos, ya que estos comportamientos límites solo dependen de la distribución de valores propios límite de las matrices deterministas del modelo. Sin embargo, existen casos en que la ESD del modelo no converge. En tales situaciones, por lo tanto, ya no hay interés en observar el comportamiento asintótico de ESD. En su lugar, estaremos interesados en encontrar equivalentes deterministicos para el modelo subyacente.

Definición 1.1.14. Considere una sucesión de matrices aleatorias $B_1, B_2, ...$ con $B_n \in \mathbb{H}_n(\mathbb{C})$. Sea $f_1, f_2, ...$ una sucesión de funcionales de $1 \times 1, 2 \times 2, ...$ matrices. Un equivalente deterministico de B_n para el funcional f_n es una sucesión $\{B_n^0\}$ con $B_n^0 \in \mathbb{M}_n(\mathbb{C})$, de matrices deterministas, tal que

$$\lim_{n\to\infty}(f_n(B_n) - f_n(B_n^0)) = 0$$

con probabilidad 1.

También, $f_n(B_n^0)$ es conocido como el equivalente deterministico de $f_n(B_n)$. A menudo se considera f_n como la traza normalizada de $Q_{B_n}(z)$, es decir, la transformada de Stieltjes de F^{B_n}.

Proposición 1.1.9. *Supóngase que $n = n(p)$, y $\frac{p}{n} \to c \in (0, \infty)$ cuando $p \to \infty$. Entonces, la matriz $m(z)I_p$ es un equivalente deterministico de $Q_{S_n}(z)$, respecto a el funcional traza normalizada.*

Demostración. Notemos que $\frac{1}{p}Tr(m(z)I_p) = m(z)$. Además por la Proposición 1.1.6, sabemos que $\frac{1}{p}Tr(Q_{S_n}(z)) - m(z)$ converge a 0 con probabilidad 1, cuando $p \to \infty$. Por lo que tomando $f_p = \frac{1}{p}Tr$, por definición de equivalente deterministico, se sigue lo deseado. □

Proposición 1.1.10. *Sea $M \in \mathbb{H}_n(\mathbb{C})$. Supongamos que $\hat{Q}_M$ es equivalente deterministico de Q_M. Entonces*
1. *Si $a, b \in \mathbb{R}^n$ son de norma euclidiana acotada,*

$$a^T(Q_M - \hat{Q}_M)b \to 0$$

con probabilidad 1.
2. *Si* $A \in \mathbb{M}_n(\mathbb{R})$ *es matriz de norma espectral acotada,*

$$\frac{1}{n}tr(AQ_M) - \frac{1}{n}tr(A\hat{Q}_M) \to 0$$

con probabilidad 1.

Como tales, los equivalentes deterministicos permiten transferir propiedades espectrales aleatorias de M en forma de cantidades limitantes deterministas y, por lo tanto, permiten una investigación más detallada. Esto haciendo un uso del equivalente deterministico en lugar de la matriz original. Es decir,

$$Q_M \leftrightarrow \hat{Q}_M.$$

1.2. Identidad para Funcionales de Matrices

En expresiones como en la Proposición 1.2.1 se estará trabajando en el Capitulo 3. El material de la Sección 1.2 puede ser encontrado en [7].

Las integrales de contorno son importantes para gran parte del análisis complejo, ya que forman el componente clave del teorema integral de Cauchy. La regla trapezoidal es importante para la aproximación de integrales debido a su precisión exponencial en ciertas circunstancias. Sin embargo, la técnica de combinar la regla trapezoidal con la integración del contorno es bastante poco referenciada o utilizada. Combinando el Teorema Integral de Cauchy y la regla de trapecio se obtiene una receta para un algoritmo muy poderoso para calcular funcionales de matrices.

Teorema 1.2.1. (Formula Integral de Cauchy para matrices)
Sea $A \in \mathbb{H}_n(\mathbb{C})$ *y* f *función analítica dentro de un contorno* γ *en* $\mathbb{C}$. *Supongamos que* γ *contiene el espectro de* A. *Entonces,*

$$f(A) = \frac{-1}{2\pi i}\oint_\gamma f(z)Q_A(z)dz.$$

Proposición 1.2.1. *Sean* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ *se sigue que*

$$\boldsymbol{a}^T f(A)\boldsymbol{b} = \frac{-1}{2\pi i}\oint_\gamma f(z)\boldsymbol{a}^T Q_A(z)\boldsymbol{b}dz.$$

1.3. Probabilidad Libre y Matrices Aleatorias

La teoría de probabilidad libre, es un enfoque diferente a las herramientas ya introducidas para la teoría de matriz aleatoria. Proporciona un marco muy eficiente para estudiar distribuciones limites de algunos modelos de matrices aleatorias de grandes dimensiones con características simétricas.

La probabilidad libre es una teoría matemática que estudia variables aleatorias no conmutativas. Esta teoría fue iniciada por Dan Voiculescu [8] alrededor de 1986 para atacar el problema del isomorfismo de factores de grupo libres, un problema importante no resuelto en la teoría de álgebras de operadores, así como suma y multiplicación de variables aleatorias no conmutativas. Más al respecto de la historia de la probabilidad libre y todo el contenido siguiente en esta sección se observa en [10], [11], [12].

1.3.1. Definiciones Básicas

Primero definimos espacios de probabilidad no-conmutativo.

Definición 1.3.1. Una Álgebra $\mathcal{A}$ sobre un campo $\mathbb{F}$, es un espacio vectorial sobre $\mathbb{F}$, que además tiene una operación de multiplicación $\cdot : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$ que es asociativa y bilineal. Si además $\mathcal{A}$ tiene un único neutro multiplicativo, se dice que $\mathcal{A}$ es unital.

Definición 1.3.2. Un espacio de probabilidad no-conmutativo es una pareja $(\mathcal{A}, \phi)$ donde $\mathcal{A}$ es una álgebra unital no-conmutativa, esto es un álgebra sobre $\mathbb{C}$ que tiene unidad denotada por 1, y $\phi : \mathcal{A} \longrightarrow \mathbb{C}$ es un funcional lineal tal que $\phi(1) = 1$.

En la definición anterior, el papel que desempeña ϕ puede ser comparado con el rol de la esperanza en probabilidad clásica.

Definición 1.3.3. Sea $(\mathcal{A}, \phi)$ un espacio de probabilidad no-conmutativo. En el contexto de probabilidad libre, una variable aleatoria es un elemento a de $\mathcal{A}$. Llamamos distribución de a a la función lineal ρ en $\mathbb{C}[X]$, el álgebra de polinomios complejos en una variable, dada por

$$\rho_a(P) = \phi(P(a)).$$

La distribución de una variable aleatoria no-conmutativa a esta caracterizada por sus momentos, los cuales están definidos como $\phi(a), \phi(a^2)$, y así sucesivamente. La distribución de una variable aleatoria no conmutativa a menudo puede asociarse con una medida de probabilidad real μ_a en el sentido de que para cada $k \in \mathbb{N}$

$$\phi(a^k) = \int_{\mathbb{R}} t^k d\mu_a(t).$$

Ejemplo 1.3.4. Consideremos el álgebra unital de matrices aleatorias de tamaño p, $\mathcal{A}_p$, Un espacio de probabilidad no-conmutativo es obtenido cuando a $\mathcal{A}_p$ le asociamos el funcional ϕ_p dado por

$$\phi_p(A) = \frac{1}{p}\mathbb{E}(Tr(A)), \quad A \in \mathcal{A}_p.$$

La distribución ρ_A, de A, es definida por el hecho de que su acción sobre cada monomio x^k de $\mathbb{C}[X]$ esta dado por,

$$\rho_A(x^k) = \phi_p(A^k).$$

Definición 1.3.5. Sea $\left\{A_p^{(1)}, ..., A_p^{(I)}\right\}$ una familia de matrices aleatorias de tamaño $p \times p$ pertenecientes al espacio de probabilidad no conmutativo $(\mathcal{A}_p, \phi_p)$. La distribución conjunta tiene una distribución limite ρ sobre $\mathbb{C}[x_i : i\,\{1, ..., I\}]$ cuando $p \to \infty$ si,

$$\rho(x_{i_1}^{k_1}...x_{i_n}^{k_n}) = \lim_{p\to\infty} \phi_p((A_p^{(1)})^{k_1}...(A_p^{i_n})^{k_n})$$

para cualquier monomio no conmutativo en $\mathbb{C}[x_i : i\,\{1, ..., I\}]$.

Cuando $m = 1$ la definición de convergencia en distribución dice que si la traza normalizada del valor esperado de A_p^k tiende a el k-ésimo momento de A, entonces A_p converge en distribución a A.

1.3.2. Independencia Libre

El concepto de independencia libre fue introducido Dan Voiculescu con el objetivo de estudiar problemas de teoría de álgebras de operadores. Notó que la relación libre se comporta de forma análoga al concepto clásico de independencia, pero en espacios de probabilidad no-conmutativos.

Definición 1.3.6. Sea $(\mathcal{A}, \phi)$ un espacio de probabilidad no-conmutativo. Una familia $\{A_1, ..., A_I\}$ de subálgebras unitales de $\mathcal{A}$ son llamadas independientes libremente (o simplemente libres) si, $\phi(a_1a_2...a_n) = 0$ para toda n-uplas $(a_1, a_2, ..., a_n)$ que satisfacen
1) $a_j \in \mathcal{A}_{i_j}$ para algún $i_j \leq I$ y $i_1 \neq i_2, i_2 \neq i_3, ..., i_{n-1} \neq i_n$.
2) $\phi(a_j) = 0$ para todo $j \in \{1, ..., n\}$.

Una familia de subconjuntos de $\mathcal{A}$ es libre si la familia de subálgebra unitales generada por cada uno de ellos es libre. Variables aleatorias $\{a_1, ..., a_n\}$ son libres si la familia de subconjuntos $\{\{a_1\}, ..., \{a_n\}\}$ es libre.

Observación. Observemos que si a_1 y a_2 son variables aleatorias libres, entonces $\phi(a_1a_2) = \phi(a_1)\phi(a_2)$. Podemos pensar la independencia libre como una fórmula para calcular momentos en conjuntos de variables aleatorias libres a partir de los momentos de las variables aleatorias individuales.

Para aplicar la teoría de la probabilidad libre a teoría de matrices aleatorias, necesitamos ampliar la definición de libre a la libertad asintótica; Es decir, reemplazando el funcional ϕ_p por el funcional

$$\varphi = \lim_{p\to\infty} \phi_p.$$

Observemos que $\varphi(I) = 1$, donde la identidad I es por definición el ensamble de matrices identidad. Entonces podemos considerar el espacio de ensambles de matrices aleatorias, $\bar{\mathcal{A}}$, dentro del marco de espacio de probabilidad libre y por lo tanto definir la relación libre para ensambles de matrices aleatorias y el funcional φ.

Definición 1.3.7. La familia $\left\{A_p^{(1)}, ..., A_p^{(I)}\right\}$ de ensambles de matrices aleatorias de tamaño $p \times p$ pertenecientes al espacio de probabilidad no conmutativo $(\bar{\mathcal{A}}, \sigma)$, es asintóticamente

libre, si satisface las siguientes dos condiciones,
1) Para cualquier entero $i \in \{1, ..., I\}$, $A_p^{(i)}$, tiene una distribución limite sobre $\mathbb{C}[X]$.
2) Para cualquier subconjunto $\{i_1, ..., i_n\} \subset \{1, ..., I\}$ con $i_1 \neq i_2, ..., i_{n-1} \neq i_n$ y para cualquier conjunto de polinomios $\{P_1, ..., P_n\}$, en una variable, satisfaciendo

$$\varphi(P_j(A_p^{(i_j)})) = 0 \quad j \in \{1, ..., n\}$$

se tiene que,

$$\varphi(\prod_{j=1}^{n} P_j(A_p^{(i_j)})) = 0.$$

Decimos que $\left\{A_p^{(1)}, ..., A_p^{(I)}\right\}$ es asintóticamente libre casi seguramente, si son asintóticamente libre excepto en un conjunto de medida cero.

La definición de libertad asintótica puede verse como libertad haciendo uso del funcional φ. Los siguientes teoremas facilitan el análisis al final del Capítulo 3, en el cual se hará uso de la teoría de probabilidad libre con el fin de otorgar rigurosidad a algunos de los pasos de la prueba del resultado principal. Estos ofrecen criterios para conocer si dos ensambles son asintóticamente libres, los cuales los se pueden ver detenidamente en [10].

Teorema 1.3.1. (Teorema de Voiculescu) *Sean $A_N^{(1)}, .., A_N^{(p)}$ matrices aleatorias gaussianas independientes de tamaño $N \times N$. Sean $D_N^{(1)}, ..., D_N^{(q)}$ matrices deterministas de tamaño $N \times N$ tales que*

$$D_N^{(1)}, ..., D_N^{(q)} \to d_1, .., d_q \text{ cuando } n \to \infty$$

en distribución. Entonces,

$$A_N^{(1)}, ..., A_N^{(p)}, D_N^{(1)}, ..., D_N^{(q)} \to s_1, ..., s_p d_1, ..., d_q \text{ cuando } n \to \infty$$

en distribución, donde cada s_i es semi-circular y $s_1, ..., s_p \{d_1, ..., d_q\}$ son libres. En particular se tiene que $A_N^{(1)}, ..., A_N^{(p)}, \left\{D_N^{(1)}, ..., D_N^{(q)}\right\}$ son asintóticamente libres.

Teorema 1.3.2. *Sea Sean $A_N^{(1)}, .., A_N^{(p)}$ matrices aleatorias gaussianas independientes de tamaño $N \times N$. Sean $D_N^{(1)}, ..., D_N^{(q)}$ matrices aleatorias de tamaño $N \times N$ tales que, casi seguramente,*

$$D_N^{(1)}, ..., D_N^{(q)} \to d_1, .., d_q \text{ cuando } n \to \infty$$

en distribución. Más aun, supóngase que $A_N^{(1)}, ..., A_N^{(p)}, \left\{D_N^{(1)}, ..., D_N^{(q)}\right\}$ son independientes. Entonces, casi seguramente

$$A_N^{(1)}, ..., A_N^{(p)}, D_N^{(1)}, ..., D_N^{(q)} \to s_1, ..., s_p d_1, ..., d_q \text{ cuando } n \to \infty$$

en distribución, donde cada s_i es semi-circular y $s_1, ..., s_p \{d_1, ..., d_q\}$ son libres. En particular se tiene que $A_N^{(1)}, ..., A_N^{(p)}, \left\{D_N^{(1)}, ..., D_N^{(q)}\right\}$ son asintóticamente libres casi seguramente.

Los Teoremas 1.3.1 y 1.3.2 pueden extenderse al caso cuando los ensambles son de matrices rectangulares de tamaño $N \times M$, agregando la hipótesis de que $\frac{M}{N} \to C$ cuando $M, N \to \infty$, donde C es una constante. Para estudiar esto último, se puede considerar la referencia [24].

1.3.3. Cumulantes

En probabilidad clásica, los momentos y los cumulantes cumplen relaciones muy útiles e interesantes, que tienen que ver con resultados combinatorios en las látices de particiones. En probabilidad libre también existen relaciones similares, pero basadas en las látices de particiones que no se cruzan. Los cumulantes libres fueron introducidos por Ronald Speicher en [11].

Definición 1.3.8. Una partición del conjunto $S = \{1, ..., n\}$ es una descomposición $\pi = \{V_1, ..., V_r\}$ de S en subconjuntos disjuntos no vacíos. Llamamos a los $V_i's$ los bloques π. Para $1 \leq p, q \leq n$ escribimos

$$p \sim_\pi q \quad \text{si p y q pertenecen al mismo bloque de } \pi.$$

La partición π es no-cruzada si lo siguiente no ocurre:
Existen $1 \leq p_1 < q_1 < p_2 < q_2 \leq n$ tales que,

$$p_1 \sim_\pi p_2 \nsim_\pi q_1 \sim_\pi q_2.$$

El conjunto de todas las particiones no-cruzadas de $\{1, ..., n\}$ se denota por $NC(n)$.

Definición 1.3.9. Sea $(\mathcal{A}, \phi)$ un espacio de probabilidad no conmutativo. Definimos el cumulante libre $\kappa_n : \mathcal{A}^n \to \mathcal{C}$ (para $n \geq 1$) como un funcional multi-lineal que satisface la relación

$$\phi(a_1, ..., a_n) = \sum_{\pi \in NC(n)} \kappa_\pi(a_1, ..., a_n)$$

donde κ_π denota un producto de cumulantes de acuerdo a la estructura a bloques de π:

$$\kappa_\pi(a_1, ..., a_n) := \kappa_{V_1}(a_1, ..., a_n)...\kappa_{V_r}(a_1, ..., a_n)$$

para $\pi = \{V_1, ..., V_r\}$ y

$$\kappa_V(a_1, ..., a_n) := \kappa_{\#V}(a_{v_1}, ..., a_{v_l}) \quad \text{para } V = (v_1, ..., v_l).$$

El siguiente teorema muestra que la libertad es mucho más fácil de describir en el nivel de los cumulantes que en el nivel de los momentos. Esta caracterización es la base de la mayoría de los cálculos con cumulantes libres.

Teorema 1.3.3. *Se tiene que,*

$$x, y \text{ son libres} \Leftrightarrow \kappa_n(a_1, ..., a_n) = 0$$

siempre que: $n \geq 2$, $a_i \in \{x, y\}$ para todo i, existe al menos dos índices i, j tales que $a_i = x, a_j = y$.

Definición 1.3.10. Sean π, σ particiones de $\{1, .., n\}$, decimos que $\pi \leq \sigma$ si cada bloque de π está contenido (como subconjunto) en algún bloque de σ, a este orden parcial se le llama refinamiento en reversa.

Con la notación de la definición anterior, decimos que

$$\phi_\sigma(a_1, ..., a_n) = \sum_{\pi \in NC(n), \pi \leq \sigma} \kappa_\pi(a_1, ..., a_n).$$

Definición 1.3.11. Sea $\pi \in NC(n)$ una partición que no se cruza de los números $1, ..., n$. Introducimos números adicionales $\bar{1}, ..., \bar{n}$, con orden alterno entre los viejos y lo nuevos. Definimos el complemento $Kr(\pi)$ como el elemento más grande (en el sentido del refinamiento en reversa) con la propiedad

$$\pi \cup \sigma \in NC(1, \bar{1}, ..., n, \bar{n}).$$

Teorema 1.3.4. *Considere $(\mathcal{A}, \phi)$ un espacio de probabilidad no conmutativo. Supóngase que $\{a_1, .., a_n\}, \{b_1, ..., b_n\}$ son libres. Entonces,*

$$\phi(a_1 b_1 a_2 b_2 ... a_n b_n) = \sum_{\pi \in NC(n)} \kappa_\pi(a_1, ..., a_n) \cdot \phi_{Kr(\pi)}(b_1, ..., b_n).$$

1.4. Distribución Infinitesimal

Como se mencionó en la sección anterior, la teoría de la probabilidad libre fue introducida por D. Voiculescu [8]. Roland Speicher [11] descubrió que la combinatoria de la teoría de probabilidad libre tiene que ver con particiones no cruzadas, en lugar del conjunto de todas las particiones el cual es usado en la teoría de probabilidad clásica. Más tarde, motivados únicamente por los aspectos combinatorios de la probabilidad libre, Biane, Goodman y Nica [15] introdujeron la teoría de probabilidad libre tipo B. La motivación original para la introducción de esta noción fue el hecho de que las particiones no cruzadas, central para la teoría de probabilidades (clásica o libre), está naturalmente asociada a los grupos simétricos, que son grupos de Weyl de grupos de Lie de tipo A. Así que sí el grupo hiperetrédico (el grupo de Weyl de un grupo de Lie de tipo B) reemplaza al grupo de simetría, se obtiene esta noción.

La libertad del tipo B se puede considerar inusual ya que no parece haber una noción obvia de positividad. Para una única variable aleatoria de tipo B, su ley se puede ver como se describe por un par de medidas (ϕ, ϕ'). Desafortunadamente, aunque se sabe que ϕ debería ser positivo, no se pudo asegurar nada acerca de ϕ'; y, de hecho, la medida ϕ' asociada a una variable semicircular tipo B no necesita ser positiva (como se observó en [16]). Belinschi y Shlyakhtenko [14] introducen la teoría de distribución infinitesimal, que es un debilitamiento de la noción de una ley de tipo B, con el fin de poder asumir la positividad en (ϕ, ϕ').

Las siguientes definiciones se obtuvieron de [13] y [14], en donde James Mingo define la distribución infinitesimal para ensambles de matrices aleatorias.

Definición 1.4.1. Una estructura $(\mathcal{A}, \phi, \phi')$ con
1. $\mathcal{A}$ una álgebra unital sobre $\mathbb{C}$
2. $\phi, \phi' : \mathcal{A} \to \mathbb{C}$ funcionales lineales tales que

$$\phi(1) = 1,\ \phi'(1) = 0$$

es llamado un espacio de probabilidad no conmutativo infinitesimal.

Definición 1.4.2. Cuando $\mathcal{A} = \mathbb{C}[X]$ en la definición anterior, el par (ϕ, ϕ') es llamado distribución infinitesimal.

Definición 1.4.3. Supóngase que $\{X_N\}_N$ es un ensamble de matrices aleatorias autoadjuntas. Si, existe
1. $\phi(x^k) = \lim_{N\to\infty} \mathbb{E}(\frac{1}{N} Tr(X_N^k))$
2. $\phi'(x^k) = \lim_{N\to\infty} N(\mathbb{E}(\frac{1}{N} Tr(X_N^k)) - \phi(x^k))$
decimos que $\{X_N\}_N$ tiene distribución infinitesimal.

La Definición 1.4.3 dice que si un ensamble de matrices aleatorias autoadjuntas tiene distribución infinitesimal, entonces por punto 1 de la Definición 1.4.3, tiene distribución límite en el sentido de probabilidad libre.

De igual forma que las otras extensiones de probabilidad libre, en distribución infinitesimal se tiene libertad infinitesimal, cumulantes infinitesimales, teoremas y aplicaciones como el presente trabajo que avalan la existencia de las distribuciones infinitesimales.

Capítulo 2

Red Neuronal Artificial

En este capítulo se introducen conceptos relacionados con redes neuronales artificiales que se emplean para desarrollar el análisis de este trabajo. Se explica el problema que se tratará en esta tesis, en el cual se utilizará una red neuronal artificial. Después, se pasa a la presentación de los supuestos con los que se trabajarán para realizar un análisis del funcionamiento de la red neuronal en el problema abordado, en el Capítulo 3.

2.1. La Red Neuronal Artificial

Una red neuronal artificial es un modelo matemático inspirado en el comportamiento biológico de las neuronas y en cómo se organizan formando la estructura del cerebro. Las redes neuronales artificiales son más que otra forma de emular ciertas características propias de los humanos, como la capacidad de memorizar y de asociar hechos. Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal artificial (donde se somete a diversas operaciones) produciendo unos valores de salida.

Las redes neuronales artificiales se han utilizado para resolver una amplia variedad de tareas, por ejemplo reconocer patrones (incluyendo imágenes, manuscritos y secuencias de tiempo) así como la visión por computador y el reconocimiento de voz, los cuales son difíciles de resolver usando enfoques clásicos.

La Sección 2.1 se basa en [18], la cual es la referencia principal de esta tesis pues los resultados que se presentan en el Capítulo 3 son el trabajo de Zhenyu Liao y Romain Couillet [18].

2.1.1. Definiciones Básicas

En la literatura, existen numerosas formas de definir a las redes neuronales artificiales; desde las definiciones cortas y genéricas hasta las que intentan explicar más detalladamente qué son las redes neuronales artificiales. En el presente trabajo se utiliza un tipo de red neuronal artificial simple. Esto debido a que Saxe, McClelland y Ganguli [17] mues-

tran empíricamente que las redes neuronales artificiales que se usan hoy en día y las redes neuronales artificiales tan simples como la que se usa en esta tesis exhiben el mismo comportamiento siguiente:

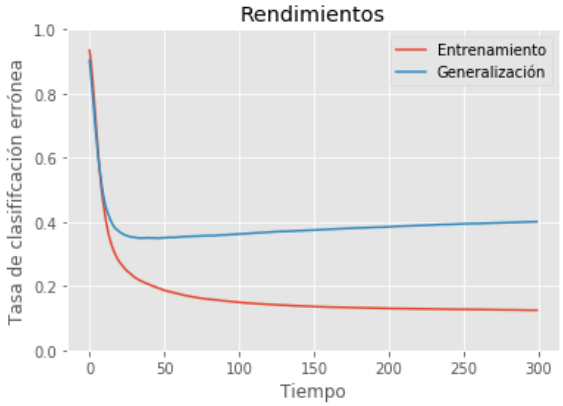

Figura 2.1: Rendimiento de la red neuronal

La curva roja, el rendimiento de entrenamiento, decae rápidamente a cero y continúa de esta manera a lo largo del tiempo. Por otro lado la curva azul, el rendimiento de generalización, decae rápidamente pero en un tiempo temprano comienza a crecer, continuando de esta manera en los tiempos sucesivos. En la Subsección 2.1.2 y 2.1.3 veremos la definición de entrenamiento y generalización respectivamente.

Sea $D \subseteq \mathbb{R}^p$ un conjunto arbitrario, llamado dominio. Cada $x \in D$ representa las características de algún objeto. Sea $\mathcal{Y} = \{-1, 1\} \subseteq \mathbb{R}$, el conjunto etiquetas. Cada $y \in \mathcal{Y}$ representa un conjunto de objetos.

Definición 2.1.1. Sea $w \in \mathbb{R}^p$. Decimos que $RNA_w : D \subseteq \mathbb{R}^p \longrightarrow B \subseteq \mathbb{R}$ es una red neuronal artificial lineal de una capa, si para $x \in D$ llamado entrada de la red; la salida de la red $RNA_w(x) = id \circ \langle w, x \rangle$ con $\langle , \rangle$ denotando al producto escalar. El vector w es conocido como vector de pesos asociado a D.

La función que en la definición anterior es la identidad, por lo general es llamada función de activación. Las funciones de activación más comunes son lineales, escalón y sigmoideas. Mientras que la función que es igual al producto escalar, es nombrada función de propagación.

Definición 2.1.2. Sea RNA_w una red neuronal artificial lineal de una capa. Para una entrada x de la red, $sign(RNA_w(x))$ es la predicción de la red con respecto a x.

Sea RNA_{w_0} una red neuronal artificial lineal de una capa con vector de pesos w_0. A continuación nos dirigimos al entrenamiento y generalización de esta red neuronal.

2.1.2. Entrenamiento

Para que la red neuronal artificial lineal de una capa empiece a predecir o "tratar" de predecir es necesario que sea previamente entrenada. Por lo visto en la Subsección 2.1.1 podemos decir que una red neuronal artificial lienal de una capa esta basada en la idea de combinar ciertos parámetros, las entradas con sus pesos, para predecir resultados por medio de la salida de la red. Por lo que llamamos entrenamiento de la red neuronal artificial al proceso de encontrar dicha combinación. Más rigurosamente,

Definición 2.1.3. Sea RNA_{w_0} una red neuronal artificial lineal de una capa con vector de pesos w_0. El entrenamiento de RNA_{w_0} es el proceso mediante el cual a partir de w_0 se obtiene un vector apropiado $w_{LS} \in \mathbb{R}^p$ para que $RNA_{w_{LS}}$, con vector de pesos w_{LS}, pueda predecir.

El término "apropiado" en la definición anterior va en el sentido de que la red neuronal artificial hace mejores predicciones usando w_{LS} que w_0, no necesariamente toda predicción debe ser correcta, pero esto es lo deseado.

Definición 2.1.4. Definimos la función de pérdida $l : \mathbb{R}^p \times D \times \mathcal{Y} \longrightarrow \mathbb{R}$ dada por

$$l(w, \hat{x}, \hat{y}) = \frac{1}{2} \left\| \hat{y} - w^T \hat{x} \right\|^2 .$$

Definición 2.1.5. Sea $\{(\hat{x}_i, \hat{y}_i)\}_{i=1}^n \in D \times \mathcal{Y}$. Definimos la función de pérdida empírica de la muestra como,

$$L(w) = \frac{\sum_{i=1}^n l(w, \hat{x}_i, \hat{y}_i)}{n} = \frac{1}{2n} \left\| \hat{Y} - w^T \hat{X} \right\|^2$$

con $\hat{X} = [\hat{x}_1, ..., \hat{x}_n]$ y $\hat{Y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_n]^T$.

Para realizar el entrenamiento de RNA_{w_0} se toman $\{(x_i, y_i)\}_{i=1}^n \in D \times \mathcal{Y}$ como datos de entrenamiento. Usando estos datos se busca minimizar el error entre la salida de la red y lo que debería salir. Esto se realiza mediante la minimización del promedio de la función de pérdida dejando fijo los datos de entrenamiento. Es decir, se minimiza la función de perdida empírica de los datos de entrenamiento.

Lema 2.1.1. *La derivada de la función de pérdida empírica* $\frac{\partial L(w)}{\partial w} = \frac{-1}{n} \hat{X}(\hat{Y} - w^T \hat{X})$.

Lema 2.1.2. *El mínimo global de la función de pérdida empírica es* $w_{LS} = (\hat{X}\hat{X}^T)^{-1}\hat{X}\hat{Y}$.

Sea $X = [x_1, ..., x_n]$ y $Y = [y_1, y_2, ..., y_n]^T$. Por Lema 2.1.2, se tiene que el vector $w_{LS} = (XX^T)^{-1}XY$ es el que permitirá a RNA_{w_0} realizar predicciones. Sin embargo, en la practica las funciones de pérdida utilizadas no son tan simples de minimizar, por lo que se hace uso de técnicas computacionales. Una de las más comunes es la conocida como descenso de gradiente. De hecho, en ocasiones esta técnica permite conocer la evolución temporal de los vectores de pesos en el entrenamiento, lo cual es mucho más informativo a la hora de realizar un análisis sobre la red neuronal artificial. Es por ello que en el presente trabajo también se considera el descenso de gradiente.

Observación. Si la función multivariable L es diferenciable en una vecindad de un punto a, entonces $L(w)$ decrece mas rápido si uno va de a en la dirección negativa del gradiente L en a. De modo que si,

$$a_{n+1} = a_n - \alpha \nabla L(a_n)$$

entonces $L(a_{n+1}) \leq L(a_n)$.

Definición 2.1.6. (Descenso de Gradiente)
Sea $\alpha \in \mathbb{R}$. Iniciamos el entrenamiento con un vector inicial $w_0 \in \mathbb{R}^p$ y se considera la sucesión $\{w_t\}_{t=0}^{\infty} \subset \mathbb{R}^p$ dada por

$$w_{t+1} = w_t - \alpha \left. \frac{\partial L(w)}{\partial w} \right|_{w=w_t}$$

para $t \geq 0$. Se espera que la convergencia de la sucesión $\{w_t\}_{t=0}^{\infty}$ proporcione el mínimo de L. Este método de encontrar el mínimo la función L es conocido como descenso de gradiente.

En [19] se puede encontrar detalladamente este algoritmo, tanto en la parte matemática como en la parte computacional.

No obstante lo anterior es un proceso computacional. De modo que si queremos ver el proceso del entrenamiento, para ser más exactos la evolución temporal de los vectores de pesos, únicamente lo podemos hacer con una aproximación a tiempo continuo.

Proposición 2.1.1. *Siguiendo la definición anterior, sea α pequeño. Supongamos que $w_t = w(t)$ para cada $t \geq 0$ y w es una función diferenciable. Entonces se tiene que,*

$$w(t) = e^{\frac{-\alpha t}{n} XX^T} w_0 + \left(I_p - e^{\frac{-\alpha t}{n} XX^T} \right) \left(XX^T \right)^{-1} XY$$

es una aproximación a la sucesión $\{w_t\}_{t=0}^{\infty}$.

Demostración. Primero observemos que mediante inducción en h se puede mostrar que

$$w(t+h) = w(t) - h\alpha \left. \frac{\partial L(w)}{\partial w} \right|_{w=w_t}.$$

De manera que,

$$\frac{w(t+h) - w(t)}{h} = -\alpha \left. \frac{\partial L(w)}{\partial w} \right|_{w=w_t}.$$

Ahora, cuando α es pequeño, w_{t+h} y w_t están cercanos uno del otro y en consecuencia $\frac{\partial w(t)}{\partial t} = -\alpha \frac{\partial L(w)}{\partial w} = \frac{\alpha}{n} X(Y - X^T w(t))$, por el Lema 2.1.1. Resolviendo este sistema de ecuaciones, se obtiene lo deseado,

$$w(t) = e^{\frac{-\alpha t}{n} XX^T} w_0 + \left(I_p - e^{\frac{-\alpha t}{n} XX^T} \right) \left(XX^T \right)^{-1} XY$$

□

La prueba detalladamente de la Proposición 2.1.1 se encuentra en [17] y [20]. Allí, se da con exactitud cada paso de la prueba y un análisis de la obtención de la Proposición 2.1.1.

2.1.3. Generalización

La generalización es una medida de la precisión con la que la red neuronal artificial puede predecir los valores de resultados para datos nunca antes vistos. Debido a que la red neuronal artificial se evalúa en muestras finitas (en el entrenamiento), la evaluación de la red neuronal artificial puede ser sensible al error de muestreo. Como resultado, las mediciones del error de predicción en los datos de entrenamiento pueden no proporcionar mucha información sobre la capacidad predictiva en los datos nuevos. El error de generalización se puede minimizar evitando el ajuste excesivo en el algoritmo de entrenamiento.

Sea $\mathbb{P} \in \mathcal{P}(D \times \mathcal{Y})$ probabilidad que explica la frecuencia de los pares (x, y). Esta $\mathbb{P}$ no es conocida, únicamente conocemos los datos de entrenamiento $\{(x_i, y_i)\}_{i=1}^{n}$ donde (x_i, y_i) son variables aleatorias independientes e idénticamente distribuidas como $\mathbb{P}$.

Definición 2.1.7. El error esperado se define como

$$L_{esp}(w) = \mathbb{E}(l(w, x, y)) = \int_{D \times E} l(w, x, y) p(x, y) dx dy$$

con $p(x, y)$ la densidad conjunta de x y y.

Definición 2.1.8. La generalización de RNA_{w_0} esta dada por $G(w) = L_{esp}(w) - L(w)$.

Comúnmente en machine learning, para los algoritmos de aprendizaje (por ejemplo una red neuronal artificial) no se considera exactamente la definición de generalización 2.1.8 (la cual es la definición estándar), dependiendo del problema en el que se esté utilizando el algoritmo. Suelen considerarse como generalización definiciones que compartan la misma fenomenológica que la Definición 2.1.8 (medir el que tan bien trabaja con nuevos datos). A continuación se presentan ejemplos donde se utilizan definiciones de generalización distintas a 2.1.8.

Ejemplo 2.1.9. $G(w) = L_{esp}(w)$

Ejemplo 2.1.10. $G(w) = \mathbb{E}(\bar{l}(w, x, y)) - L(w)$ con $\bar{l}$ función de perdida distinta a l.

Ejemplo 2.1.11. $G(w) = \mathbb{E}(\bar{l}(w, x, y))$ con $\bar{l}$ función de perdida distinta a l.

En el Capítulo 3, cuando se trabaje con la generalización de la red neuronal artificial se estará utilizando la definición del Ejemplo 2.1.11 con $\bar{l}(w, x, y) = 1_{[sign(RNA(x)) \neq y]}$.

2.2. El Problema

Los problemas de clasificación binaria es un tema central en el aprendizaje automático que tienen que ver con con la tarea de clasificar los elementos de un conjunto dado en dos grupos (predecir a qué grupo pertenece cada uno) sobre la base de una regla de clasificación.

Consideremos RNA_{w_0} una red neuronal artificial lineal de una capa con vector de pesos w_0 y dominio el conjunto

$$D = \{x = y\mu + z : y \in \{1, -1\}, z \sim \mathcal{N}(0_p, I_p)\}$$

con μ vector fijo en $\mathbb{R}^p$, 0_p el vector en $\mathbb{R}^p$ con entradas igual a cero y I_p la matriz identidad de tamaño p.
En el contexto de un problema de clasificación binaria, RNA_{w_0} trata de predecir elementos $x \in D$ con etiqueta $y \in \mathcal{Y}$, la cual es la regla de clasificación.

Cabe mencionar que x tiene distribución $\mathcal{N}(-\mu, I_p)$ o $\mathcal{N}(\mu, I_p)$ dependiendo de si la etiqueta $y = -1$ o $y = 1$, respectivamente, es decir, dependiendo de la clase a cual pertenece x. La etiquetas y pueden verse como una variable aleatoria que tienen distribución Bernoulli, $Ber_{\pm 1}(r)$ con $r > 0$, independiente de z. Adicionalmente la independencia de los vectores aleatorios en cada clase proporciona independencia entre los vectores gausianos estándar correspondientes. RNA_{w_0} será la red neuronal artificial que se analizará.

Recordando la Sección 2.1 esta red neuronal artificial debe ser entrenada, haciendo uso de $\{x_1, ..., x_n\}$ vectores extraídos del dominio D, independientes, con sus respectivas etiquetas $\{y_1, y_2, ..., y_n\}$, $y_i \in \mathcal{Y}$. Sea l la función de pérdida y L la función de pérdida empírica como en la Definición 2.1.4 y 2.1.5 respectivamente. Consideremos $X = [x_1, ..., x_n]$ y $Y = [y_1, ..., y_n]^T$. Por el Lema 2.1.2, $w_{LS} = (XX^T)^{-1}XY$ es el que permitirá a RNA_{w_0} realizar predicciones, además por la Proposición 2.1.1,

$$w(t) = e^{\frac{-\alpha t}{n}XX^T} w_0 + \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) w_{LS}$$

es aproximación a la evolución temporal del entrenamiento. Cuando $t \to \infty$, $w(t) \to w_{LS}$ por lo que la red neuronal artificial olvida la inicialización w_0 y da como resultado el mínimo global de L. También, podemos reescribir

$$w(t) = Ue^{-\alpha t\Lambda}U^* w_0 + \left(I_p - Ue^{-\alpha t\Lambda}U^*\right) w_{LS}$$

donde $\frac{1}{n}XX^T = U\Lambda U^*$ con U matriz unitaria y Λ matriz diagonal formada por los valores propios de $\frac{1}{n}XX^T$. De manera que el núcleo de este estudio es la comprensión de los valores y vectores propios de la matriz de covarianza muestral de los datos de entrenamiento, la cual ha sido ampliamente estudiada en la literatura de matrices aleatorias. Como se vio en el Capítulo 1, específicamente Teorema 1.1.2, Proposiciones 1.1.6, 1.1.8 y 1.2.1, se tiene una amplia gama de herramientas para realizar un análisis detallado de la red neuronal artificial RNA_{w_0}. Este análisis se ve reflejado en el próximo capítulo, pero antes de proceder a esté necesitaremos trabajar bajo ciertos supuestos, con el fin de poder utilizar las herramientas mencionadas.

Observación. Podemos reescribir el vector w_{LS} como $W_{LS} = \left(XX^T\right)^{-1} XY = \left(\frac{1}{n}XX^T\right)^{-1} \frac{1}{n}XY$.

2.3. Los Supuestos

Se tiene un buen análisis del entrenamiento de la red neuronal artificial en el problema dado. Así que se continúa con el análisis del rendimiento de rendimiento de generalización. Para ello se trabaja con los siguientes supuestos:
1. Sea $p = p(n)$, $\frac{p}{n} \to c \in (0, \infty)$; cuando n $\to \infty$.
2. $||\mu|| = O(1)$.
3. w_0 es un vector aleatorio con entradas i.i.d de media cero, varianza σ^2/p.

Los primeros dos supuestos aseguran que la matriz de Wishart de grado n, normalizada, $\frac{1}{n}XX^T$, tiene norma de operador acotada para todo n, p grande con probabilidad 1. Este fue un trabajado realizado por Bai y Silverstein [4] en 1998.

2.4. Notas Adicionales

En la Sección 2.2 surgieron de manera natural las matrices aleatorias, para el entrenamiento de la red neuronal artificial. Puntualmente, se vio que la evolución temporal del entrenamiento de la red neuronal artificial depende únicamente de la matriz de convarianza muestral de los datos de entrenamiento.

Se vuelve hacer énfasis de que el trabajo que se presenta a continuación fue realizado por Couillet y su estudiante de doctorado Liao en [18]. Esto ya que al final del Capítulo 3 se presenta una replica de una parte de este trabajo, la cual se logró en esta tesis, con el fin de dar más rigurosidad a las pruebas de Couillet y Liao.

Las referencias para la Subsección 2.1.3 son [21] y [22].

Capítulo 3

Comportamiento Asintótico de la Red Neuronal Artificial

En este capítulo se presenta un análisis de la dinámica de generalización de la red neuronal artificial lineal RNA_{w_0} que se mostró en el Capítulo 2, en el problema también establecido en el Capítulo 2. Este análisis fue realizado por Liao y Couillet [18]. En la Sección 3.1 aparece el teorema para el rendimiento de generalización de la red neuronal artificial que resultó del trabajo de [18]. También se da la demostración de este teorema, a diferencia que en [18], en esta ocasión a detalle haciendo uso de las herramientas vistas en el Capítulo 1 sobre matrices aleatorias. En esta parte, el concepto de equivalente deterministico jugará un papel fundamental. Posteriormente se exponen algunos de los pasos de la demostración del teorema en que se hace uso de equivalentes deterministicos, empleando la teoría de distribución infinitesimal y probabilidad libre con el fin de proveer mayor rigurosidad a la prueba.

3.1. Rendimiento de Generalización Vía Equivalentes Deterministicos

En esta sección se efectúa un análisis del rendimiento de generalización de RNA. Específicamente observando el comportamiento asintótico de la transformada de Cauchy de la ESD de la matriz $\frac{1}{n}XX^T$, mediante una reescritura de esta matriz en términos de una normalización de la matriz de Wishart con n grados de libertad, para la cual, como se observa en las Subecciones 1.1.3, 1.1.4 y Sección 1.2 existen herramientas basadas en la teoría de matrices aleatorias para realizar un análisis de este tipo de matrices.

3.1.1. EL Resultado Principal

A continuación se presenta el teorema principal de este capítulo, para el cual, su demostración es seguida después de un conjunto de pasos en los que poco a poco se ira desenvolviendo cada término que aparece en este teorema.

Recordemos que estamos considerando RNA_{w_0} una red neuronal artificial lineal de una capa con vector de pesos w_0 y dominio el conjunto

$$D = \{x = y\mu + z : y \in \mathcal{Y} := \{1, -1\}, z \sim \mathcal{N}(0_p, I_p)\}$$

con μ vector fijo en $\mathbb{R}^p$, 0_p el vector en $\mathbb{R}^p$ con entradas igual a cero, I_p la matriz identidad de tamaño p. Además RNA_{w_0} es entrenada con $\{x_1, ..., x_n\}$ vectores extraídos del dominio D, independientes, con sus respectivas etiquetas $\{y_1, y_2, ..., y_n\}$, $y_i \in \mathcal{Y}$. Se Considera $X = [x_1, ..., x_n]$ y $Y = [y_1, ..., y_n]^T$.

Teorema 3.1.1. *Sea $(x, y) \in D \times \mathcal{Y}$, no dato de entrenamiento. Cuando $n \to \infty$; con probabilidad* 1

$$P\left(w(t)^T x > 0 \mid y = -1\right) - Q\left(E/\sqrt{V}\right) \to 0$$

$$P\left(w(t)^T x < 0 \mid y = 1\right) - Q\left(E/\sqrt{V}\right) \to 0$$

donde:

$$E = \frac{-1}{2i\pi} \oint_\gamma \frac{1 - f_t(z)}{z} \frac{||\mu||^2 m(z)}{(||\mu||^2 + c)m(z) + 1} dz$$

$$V = \frac{1}{2i\pi} \oint_\gamma \left(\frac{\frac{1}{z^2}(1 - f_t(z))^2}{(||\mu||^2 + c)m(z) + 1} - \sigma^2 (f_t(z))^2 m(z) \right) dz$$

con $f_t(z) = exp(-\alpha t z)$, $m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{(1-c-z)^2 - 4cz}}{2cz}$ la transformada de stieltjes de la distribución Marchenko-Pastur (Proposición 1.1.4), $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{u^2}{2})\, \mathrm{d}u$ la función conocida como Q-función y γ un camino cerrado orientado positivamente que contiene todos los valores propios de $\frac{1}{n} X X^T$ y al origen.

Recordando la definición de la generalización en el Ejemplo 2.1.11, el Teorema 3.1.1 manifiesta que el rendimiento de generalización aunque es aleatorio tiene un comportamiento asintóticamente deterministico descrito por el par (E, V). Es aun más provechoso que el par (E, V) depende únicamente de los valores de σ, α, μ y c, los cuales, como se vio en el Capítulo 2, son parámetros de entrada "indirectamente" a $RNAw_0$. Debido a lo cual, podemos mejorar o empeorar el rendimiento de generalización de RNA_{w_0} con la elección de estos valores. Por ejemplo [23] demostró que $\sigma = 0$ es perjudicial para el rendimiento de la una red neuronal.

3.1.2. La Demostración

Para la demostración del Teorema 3.1.1 se verá en seguida una serie de lemas y proposiciones que serán de ayuda.

Proposición 3.1.1. *Sea $x \in D$ con etiqueta y, entonces*

$$P\left(w(t)^T x > 0 \mid y = -1\right) = Q\left(\frac{\mu^T w(t)}{||w(t)||}\right)$$

$$P\left(w(t)^T x < 0 \mid y = 1\right) = Q\left(\frac{\mu^T w(t)}{||w(t)||}\right).$$

Demostración. Primero nos percatamos que el vector aleatorio x es independiente de $w(t)$. En consecuencia, $w(t)^T x$ es un vector gaussiano de media $-\mu^T w(t)$ y varianza $w(t)^T I_p w(t) = ||w(t)||^2$ para $y = -1$, y de media $\mu^T w(t)$ y varianza $||w(t)||^2$ si $y = 1$. Ahora, como $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{u^2}{2})\, du$ es la función de distribución de la cola de la distribución normal estándar. Se sigue que:

1. Si $y = -1$. Entonces,

$$\begin{aligned} P\left(w(t)^T x > 0 \mid y = -1\right) &= Q\left(\frac{0 - \mathbb{E}(w(t)^T x)}{\sqrt{Var(w(t)^T x)}}\right) \\ &= Q\left(\frac{\mu^T w(t)}{||w(t)||}\right). \end{aligned} \tag{3.1}$$

2. Si $y = 1$. Entonces,

$$\begin{aligned} P\left(w(t)^T x < 0 \mid y = 1\right) &= Q\left(\frac{0 - \mathbb{E}(-w(t)^T x)}{\sqrt{Var(-w(t)^T x)}}\right) \\ &= Q\left(\frac{\mu^T w(t)}{||w(t)||}\right). \end{aligned} \tag{3.2}$$

De aquí se obtiene lo deseado. □

La Proposición 3.1.1 provee un camino para conocer el comportamiento asintótico del rendimiento de generalización mediante el comportamiento asintótico de $\mu^T w(t)$ y $||w(t)||$. No obstante estas dos últimas expresiones necesitan ser manipuladas para percibir dicho comportamiento, lo cual se realiza empleando los siguientes lemas.

Lema 3.1.2. *Podemos escribir $\mu^T w(t)$ de la siguiente manera,*

$$\mu^T w(t) = -\frac{1}{2i\pi} \oint_\gamma f_t(z) \mu^T \left(\frac{1}{n} X X^T - z I_p\right)^{-1} w_0 dz$$

$$-\frac{1}{2i\pi} \oint_\gamma \frac{(1 - f_t(z))}{z} \mu^T \left(\frac{1}{n} X X^T - z I_p\right)^{-1} \frac{1}{n} X Y dz.$$

Demostración. Tenemos que, por Proposición 2.1.1,

$$\mu^T w(t) = \mu^T e^{\frac{-\alpha t}{n}XX^T} w_0 + \mu^T \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) w_{LS}$$

$$= \mu^T e^{\frac{-\alpha t}{n}XX^T} w_0 + \mu^T \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) \left(\frac{1}{n}XX^T\right)^{-1} \frac{1}{n}XY$$

donde la última igualdad es debido a la observación de la Sección 2.2. Luego, por fórmula integral de Cauchy para matrices Teorema 1.2.1 y Proposición 1.2.1, se obtiene que:

1. Tomando $f_0(z) = e^{-\alpha t z} =: f_t(z)$ resulta que,

$$\mu^T e^{\frac{-\alpha t}{n}XX^T} w_0 = -\frac{1}{2i\pi} \oint_\gamma f_t(z) \mu^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} w_0 dz. \tag{3.3}$$

2. Tomando $f_1(z) = (1 - e^{-\alpha t z})z^{-1} = (1 - f_t(z))z^{-1}$ resulta que

$$\mu^T \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) w_{LS} = -\frac{1}{2i\pi} \oint_\gamma \frac{(1 - f_t(z))}{z} \mu^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} \frac{1}{n}XY dz \tag{3.4}$$

con γ un camino cerrado orientado positivamente que contiene a los valores propios de $\frac{1}{n}XX^T$. Finalmente de las ecuaciones (3.3) y (3.4) se obtiene el lema. □

Lema 3.1.3. *Podemos escribir $w(t)^T w(t)$ de la siguiente manera,*

$$w(t)^T w(t) = -\frac{1}{2i\pi} \oint_\gamma f_t^2(z) w_0^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} w_0 dz$$

$$-\frac{1}{i\pi} \oint_\gamma \frac{f_t(z)(1 - f_t(z))}{z} w_0^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} \frac{1}{n}XY dz$$

$$-\frac{1}{2i\pi} \oint_\gamma \frac{(1 - f_t(z))^2}{z^2} \frac{1}{n} Y^T X^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} \frac{1}{n}XY dz.$$

Demostración. Tenemos que, por la Proposición 2.1.1 (la expresión de $w(t)$)

$$w(t)^T w(t) = w_0^T \left(e^{\frac{-\alpha t}{n}XX^T}\right)^2 w_0 + w_{LS}^T e^{\frac{-\alpha t}{n}XX^T} \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) w_0$$

$$+ w_0^T e^{\frac{-\alpha t}{n}XX^T} \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) w_{LS} + w_{LS}^T \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right)^2 w_{LS}$$

$$= w_0^T \left(e^{\frac{-\alpha t}{n}XX^T}\right)^2 w_0 + \left(\left(\frac{1}{n}XX^T\right)^{-1} \frac{1}{n}XY\right)^T e^{\frac{-\alpha t}{n}XX^T} \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) w_0$$

$$+ w_0^T e^{\frac{-\alpha t}{n}XX^T} \left(I_p - e^{\frac{-\alpha t}{n}XX^T}\right) \left(\frac{1}{n}XX^T\right)^{-1} \frac{1}{n}XY$$

$$+\left(\left(\frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XY\right)^T\left(I_p-e^{\frac{-\alpha t}{n}XX^T}\right)^2\left(\frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XY.$$

donde la última igualdad se da por la observación al final del Capítulo 2. Haciendo uso del Teorema 1.2.1 y la Proposición 1.2.1 (la formula integral de Cauchy para matrices) se obtiene que

1. Tomando $f_0(z)=(e^{-\alpha tz})^2=f_t(z)^2$, resulta que

$$w_0^T e^{\frac{-2\alpha t}{n}XX^T}w_0=-\frac{1}{2i\pi}\oint_\gamma f_t^2(z)w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}w_0dz. \tag{3.5}$$

2. Tomando $f_1(z)=e^{-\alpha tz}(1-e^{-\alpha tz})z^{-1}=f_t(z)(1-f_t(z))z^{-1}$, resulta que

$$\left(\left(\frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XY\right)^T\left(e^{\frac{-\alpha t}{n}XX^T}-e^{\frac{-2\alpha t}{n}XX^T}\right)w_0$$

$$=-\frac{1}{2i\pi}\oint_\gamma\frac{f_t(z)(1-f_t(z))}{z}\left(\frac{1}{n}XY\right)^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}w_0dz \tag{3.6}$$

y además,

$$w_0^T\left(e^{\frac{-\alpha t}{n}XX^T}-e^{\frac{-2\alpha t}{n}XX^T}\right)\left(\frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XY$$

$$=-\frac{1}{2i\pi}\oint_\gamma\frac{f_t(z)(1-f_t(z))}{z}w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XYdz. \tag{3.7}$$

3. Tomando $f_2(z)=(1-e^{-\alpha tz})^2(z^{-2})=(1-f_t(z))^2(z^{-2})$, resulta que

$$\left(\left(\frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XY\right)^T e^{\frac{-2\alpha t}{n}XX^T}\left(\frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XY$$

$$=-\frac{1}{2i\pi}\oint_\gamma\frac{(1-f_t(z))^2}{z^2}\frac{1}{n}Y^TX^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XYdz. \tag{3.8}$$

Finalmente tras darse cuenta que las ecuaciones (3.6) y (3.7) son la misma, se suman (3.5), (3.6), (3.7) y (3.8) para conseguir el lema. □

Por medio de Lema 3.1.2 y 3.1.3 reducimos el objetivo de conocer el límite de $\mu^T w(t)$ y $||w(t)||$ a conocer el comportamiento asintótico de expresiones de la forma

$$\mathbf{a}^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\mathbf{b}=\mathbf{a}^TQ_{\frac{1}{n}XX^T}(z)\mathbf{b},\ \text{ para }\ \mathbf{a},\mathbf{b}=\mu,w_0,\frac{1}{n}XY. \tag{3.9}$$

Para ello, reescribimos la matriz de los datos de entrenamiento de la siguiente manera.

Sean $z_1, .., z_n$ los vectores gaussianos estandar correspondientes a los vectores de entrenamiento. Entonces si $Z = [z_1, .., z_n]$, se sigue que

$$\begin{aligned} X &= [x_1, ..., x_n] \\ &= [(-1)^{a_1}\mu + z_1, ..., (-1)^{a_n}\mu + z_n] \\ &= [(-1)^{a_1}\mu, ..., (-1)^{a_n}\mu] + [z_1, ..., z_n] \\ &= [(-1)^{a_1}\mu, ..., (-1)^{a_n}\mu] + Z \end{aligned}$$

además,

$$\begin{aligned} [(-1)^{a_1}\mu, ..., (-1)^{a_n}\mu] &= \begin{bmatrix} (-1)^{a_1}\mu_1 & (-1)^{a_2}\mu_1 & \dots (-1)^{a_n}\mu_1 \\ (-1)^{a_1}\mu_2 & (-1)^{a_2}\mu_2 & \dots (-1)^{a_n}\mu_2 \\ \vdots & \vdots & \vdots \\ (-1)^{a_1}\mu_p & (-1)^{a_2}\mu_p & \dots (-1)^{a_n}\mu_p \end{bmatrix} \text{ donde } \mu = [\mu_1, ..., \mu_p]^T \\ &= \mu Y^T. \end{aligned}$$

De manera que, $X = \mu Y^T + Z$. Esto, como veremos en el Lema 3.1.4, permite hacer uso del resolvente de la matriz wishart de n grados de libertad normalizada $\frac{1}{n}ZZ^T$ para encontrar una formula del resolvente de la matriz de covarianza muestral de los datos de entrenamiento. En adelante, se tendrá en cuenta que $Q(z) := Q_{\frac{1}{n}ZZ^T}(z)$.

Lema 3.1.4. *El resolvente de la matriz de covarianza muestral de los datos de entrenamiento de RNA_{w_0} es igual a*

$$Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix} \mu^T Q(z)\mu & 1 + \frac{1}{n}\mu^T Q(z) ZY \\ 1 + \frac{1}{n}Y^T Z^T Q(z)\mu & -1 + \frac{1}{n}Y^T Z^T Q(z)\frac{1}{n}ZY \end{bmatrix}^{-1}\begin{bmatrix} \mu^T \\ \frac{1}{n}Y^T Z^T \end{bmatrix} Q(z)$$

Demostración. Sabemos que $X = \mu Y^T + Z$. De tal forma que,

$$\begin{aligned} \frac{1}{n}XX^T &= \frac{1}{n}(\mu Y^T + Z)(Y\mu^T + Z^T) \\ &= \frac{1}{n}ZZ^T + \mu\mu^T + \frac{1}{n}ZY\mu^T + \mu\frac{1}{n}Y^T Z^T \\ &= \frac{1}{n}ZZ^T + \begin{bmatrix}\mu + \frac{1}{n}ZY & \mu\end{bmatrix}\begin{bmatrix}\mu^T \\ \frac{1}{n}Y^T Z^T\end{bmatrix} \\ &= \frac{1}{n}ZZ^T + \begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix}1 & 1 \\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu^T \\ \frac{1}{n}Y^T Z^T\end{bmatrix} \end{aligned}$$

y en consecuencia,

$$\left(\frac{1}{n}XX^T - zI_p\right)^{-1} = \left[\left(\frac{1}{n}ZZ^T - zI_p\right) + \begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix}1 & 1 \\ 1 & 0\end{bmatrix}\begin{bmatrix}\mu^T \\ \frac{1}{n}Y^T Z^T\end{bmatrix}\right]^{-1}$$

Ahora, puesto que la matriz $\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}$ es invertible, hacemos uso de la identidad de Woodbury para matrices (del Apéndice B), obteniendo que

$$\left(\frac{1}{n}XX^T - zI_p\right)^{-1} = Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\left(\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}^{-1} + \begin{bmatrix}\mu^T\\ \frac{1}{n}Y^TZ^T\end{bmatrix}Q(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\right)^{-1}\begin{bmatrix}\mu^T\\ \frac{1}{n}Y^TZ^T\end{bmatrix}Q(z)$$

$$= Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\left(\begin{bmatrix}0 & 1\\ 1 & -1\end{bmatrix} + \begin{bmatrix}\mu^TQ(z)\mu & \frac{1}{n}\mu^TQ(z)ZY\\ \frac{1}{n}Y^TZ^TQ(z)\mu & \frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}\right)^{-1}\begin{bmatrix}\mu^T\\ \frac{1}{n}Y^TZ^T\end{bmatrix}Q(z)$$

$$= Q(z) - Q(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix}\mu^TQ(z)\mu & 1+\frac{1}{n}\mu^TQ(z)ZY\\ 1+\frac{1}{n}Y^TZ^TQ(z)\mu & -1+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}^{-1}\begin{bmatrix}\mu^T\\ \frac{1}{n}Y^TZ^T\end{bmatrix}Q(z).$$

De aquí se concluye el resultado. □

El Lema 3.1.4, aunque es simple, es escencial para la realización del análisis. A causa del Lema 3.1.4, cualquier operación $\mathbf{a}^TQ_{\frac{1}{n}XX^T(z)}\mathbf{b}$, se puede ver en terminos de $\mathbf{a}^TQ(z)\mathbf{b}$ para el cual, haciendo uso de equivalentes deterministicos, concretamente por la Proposiciónes 1.1.8 y la Subsección 1.1.3, podemos identificar su comportamiento asintótico.

Más aun, ya que $X = \mu Y + Z$ y tenemos la ecuación (3.9), es suficiente conocer el comportamiento asintótico de las siguientes expresiones:

$$\mathbf{a}^TQ(z)\mathbf{b}, \text{ para } \mathbf{a},\mathbf{b} = \mu, w_0, \frac{1}{n}ZY. \tag{3.10}$$

Proposición 3.1.2. ***1***. *Se tiene que* $\mu^TQ(z)\mu = m(z)||\mu||^2$

2. *Se tiene que* $\mu^TQ(z)w_0 \to 0$ *con probabilidad 1, cuando* $n \to \infty$.

3. *Se tiene que* $\frac{1}{n}\mu^TQ(z)ZY \to 0$ *con probabilidad 1, cuando* $n \to \infty$.

4. *Se tiene que* $\frac{1}{n^2}Y^TZ^TQ(z)ZY$ *converge con probabilidad 1, cuando* $n \to \infty$.

5. *Se tiene que* $\frac{1}{n}w_0^TQ(z)ZY \to 0$ *con probabilidad 1, cuando* $n \to \infty$.

6. *Se tiene que* $w_0^TQ(z)w_0 \to \sigma^2m(z)$ *con probabilidad 1, cuando* $n \to \infty$.

Demostración. Por la Propisición 1.1.9, la matriz $m(z)I_p$ es un determinista equivalente de la matriz $Q(z)$. Por este motivo, disponemos de $m(z)I_p \leftrightarrow Q(z)$.
1. Consecuentemente,

$$\mu^TQ(z)\mu = \mu^Tm(z)I_p\mu = m(z)\mu^T\mu = m(z)||\mu||^2.$$

2. De igual forma, $\mu^TQ(z)w_0 = m(z)\mu^Tw_0$. Puesto que $m(z)\mu$ es un vector en $\mathbb{R}^p$ deterministico y, por el supuesto 3, $w_0 \sim \mathcal{N}(0_p, \frac{\sigma^2}{p}I_p)$, se tiene que $m(z)\mu^Tw_0 \sim \mathcal{N}(0_p, m(z)^2\frac{\sigma^2}{p}||\mu||^2)$.

Por lo que, usando la desigualdad de Mill (Apéndice B),

$$\mathbb{P}(|\mu^T Q(z) w_0| > \epsilon) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{\left(\frac{-\epsilon^2}{2m(z)^2 \frac{\sigma^2}{p} ||\mu||^2}\right)}}{\epsilon}.$$

Por otro lado, por supuesto 1 se tiene que $||\mu||^2 = O(1)$, de donde $||\mu||^2 \leq C$ para todo p, con C constante. Se sigue entonces que,

$$\frac{1}{\sqrt{2\pi}} \frac{e^{\left(\frac{-\epsilon^2}{2m(z)^2 \frac{\sigma^2}{p} ||\mu||^2}\right)}}{\epsilon} \leq \frac{1}{\sqrt{2\pi}} \frac{e^{\left(\frac{-p\epsilon^2}{2m(z)^2 \sigma^2 C}\right)}}{\epsilon}.$$

Luego, la serie $\sum_{p=1}^{\infty} e^{\left(\frac{-p\epsilon^2}{2m(z)^2\sigma^2 C}\right)}$ es convergente pues, usando desigualdades de la función exponencial (Proposición B.0.3 del Apéndice B)

$$e^{\left(\frac{-p\epsilon^2}{2m(z)^2\sigma^2 C}\right)} \leq \frac{2}{\frac{p^2\epsilon^4}{2m(z)^4\sigma^4 C^2}} \text{ para todo } p \in \mathbb{N}$$

y la serie $\sum_{p=1}^{\infty} \frac{2}{\frac{p^2\epsilon^4}{2m(z)^4\sigma^4 C^2}}$ es convergente. De lo anterior, $\sum_{p=1}^{\infty} \mathbb{P}(|\mu^T Q(z) w_0| > \epsilon)$ es convergente. Como cuando $n \to \infty$ se tiene que $p \to \infty$, por la Proposición A.0.1 (Apéndice A), se concluye que $\mu^T Q(z) w_0$ converge a 0 con probabilidad 1, cuando $n \to \infty$.

3. Nuevamente, usando el equivalente deterministico, $\frac{1}{n}\mu^T Q(z) ZY = \frac{1}{n} m(z) \mu^T ZY$. Como el vector Y tiene entradas igual a $\pm 1's$ y además es independiente de Z, donde las entradas de la matriz Z son iid normales estandar, resulta que ZY es un vector con entradas independientes distribuidas como $\mathcal{N}(0, n)$. Entonces $\mu^T ZY$ es normal tal que

$$\mathbb{E}(\mu^T ZY) = 0, Var(\mu^T ZY) = n||\mu||^2$$

lo cual implica $\frac{1}{n} m(z) \mu^T ZY \sim \mathcal{N}(0, \frac{m(z)^2 ||\mu||^2}{n})$. De modo que, por la misma desigualdad que en (**2**), la desigualdad de Mill (Apéndice B),

$$\mathbb{P}(\frac{1}{n} m(z) |\mu^T Q(z) ZY| > \epsilon) \leq \frac{1}{\sqrt{2\pi}} \frac{e^{\left(\frac{-\epsilon^2}{2\frac{m(z)^2||\mu||^2}{n}}\right)}}{\epsilon} = \frac{1}{\sqrt{2\pi}} \frac{e^{\left(\frac{-n\epsilon^2}{2m(z)^2||\mu||^2}\right)}}{\epsilon}.$$

De aquí, siguiendo la misma linea de argumentos que en (**2**), concluimos que la serie

$$\sum_{n=1}^{\infty} \mathbb{P}(\frac{1}{n} |\mu^T Q(z) ZY| > \epsilon) < \infty$$

y por lo tanto $\frac{1}{n}\mu^T Q(z) ZY$ converge a cero con probabilidad 1.

4. Sea $\tilde{Q}(z) = \left(\frac{1}{n}Z^TZ - zI_n\right)^{-1}$ el co-resolvente de la matriz wishart normalizada $\frac{1}{n}ZZ^T$. Recordemos que en la Proposición 1.1.9, se obtuvo que $\frac{1}{n}Tr\tilde{Q}(z)$ converge con probabilidad 1, cuando $n \to \infty$, a $\tilde{m}(z)$ la cual cumple la ecuación

$$\tilde{m}(z) = cm(z) + \frac{1}{z}(c-1). \tag{3.11}$$

Ahora observemos que, por la relación de conmutatividad que cumplen el resolvente y co-resolvente

$$\begin{aligned}
\frac{1}{n^2}Y^TZ^TQ(z)ZY &= \frac{1}{n^2}Y^T\tilde{Q}(z)Z^TZY \\
&= \frac{1}{n}Y^T\tilde{Q}(z)\left(\frac{1}{n}Z^TZ - zI_n + zI_n\right)Y \\
&= \frac{1}{n}Y^T\tilde{Q}(z)\left(\frac{1}{n}Z^TZ - zI_n\right)Y + \frac{1}{n}Y^T\tilde{Q}(z)\left(zI_n\right)Y \\
&= \frac{1}{n}Y^T\left(\frac{1}{n}Z^TZ - zI_n\right)^{-1}\left(\frac{1}{n}Z^TZ - zI_n\right)Y + \frac{1}{n}Y^T\tilde{Q}(z)\left(zI_n\right)Y \\
&= \frac{1}{n}||Y^T||^2 + z\frac{1}{n}Y^T\tilde{Q}(z)Y \\
&= 1 + z\frac{1}{n}Tr\tilde{Q}(z).
\end{aligned}$$

De modo que, $\frac{1}{n^2}Y^TZ^TQ(z)ZY \to 1 + z\tilde{m}(z)$ con probabilidad 1, cuando $n \to \infty$.

5. Otra vez, por el equivalente deterministico de $Q(z)$, $w_0^TQ(z)w_0 = m(z)||w_0||^2$. Así que basta análizar la convergencia de la norma al cuadrado del vector w_0. Para ello, notemos que

$$||w_0||^2 = w_{0_1}^2 + ... + w_{0_p}^2 \text{ con } w_{0_i} \text{ la entrada } i - \text{esima entrada.}$$

Luego, $\mathbb{E}(w_{0_i}^2) = \frac{\sigma^2}{p}$, de donde $\mathbb{E}(pw_{0_i}^2) = \sigma^2$, pues cada entrada del vector w_0 tiene media cero y varianza $\frac{\sigma^2}{p}$. Más aun estas entradas son independientes y por lo tanto sus cuadrados también lo son, lo cual implica, por la ley de grandes números (el Teorema A.0.3), con probabilidad 1

$$\frac{p(w_{0_1}^2 + ... + w_{0_p}^2)}{p} \to \sigma^2$$

esto es,

$$||w_0||^2 \to \sigma^2.$$

Nuevamente recordemos que $n \to \infty$ implica $p \to \infty$, para concluir el resultado.

6. Una vez más hacemos uso del equivalente deterministico para obtener que $\frac{1}{n}w_0^TQ(z)ZY = \frac{1}{n}m(z)w_0^TZY$. Después por lo visto en (**3**) se tiene que $ZY_i \sim \mathcal{N}(0, n)$ para $i = 1, ...p$ representando la entrada i-esima del vector. Además las entradas son independientes. Por

lo que, del hecho de que w_0 es independiente de ZY, resulta que la variable aleatoria $\frac{1}{n}m(z)w_0^T ZY|w_0$ (es decir, la variable aleatoria condicionada a w_0) es una variable aleatoria gaussiana con media 0 y varianza $\frac{m(z)^2||w_0||^2}{n}$. Luego,

$$\mathbb{P}(|\frac{1}{n}m(z)w_0^T ZY| > \epsilon) = \mathbb{E}(\mathbb{P}(|\frac{1}{n}m(z)w_0^T ZY| > \epsilon|w_0)) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{\left(\frac{n\epsilon^2}{2m(z)^2||w_0||^2}\right)}}{\epsilon}$$

y como por (**5**) se tiene que $||w_0||^2$ es convergente y en consecuencia de $O(1)$, bajo la misma linea de argumentos que (**2**) y (**3**) obtenemos lo deseado. □

Los siguientes lemas que presentamos muestran el comportamiento asintótico en probabilidad, de las expresiones de la forma $a^T Q_{\frac{1}{n}XX^T}(z)b$ involucradas en los Lemas 3.1.2 y 3.1.3 y de esta manera estar listos preparados para escribir la demostración de el teorema.

Lema 3.1.5. *Cuando $n \to \infty$, con probabilidad 1*

$$\mu^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} \frac{1}{n}XY \to \frac{||\mu||^2 m(z)}{m(z)(c + ||\mu||^2) + 1}.$$

Demostración. Por el Lema 3.1.4, $\mu^T(\frac{1}{n}XX^T - zI_p)^{-1}\frac{1}{n}XY$ es igual a

$$\mu^T Q(z)\left(\frac{1}{n}XY\right) - \mu^T Q(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix}\mu^T Q(z)\mu & 1 + \frac{1}{n}\mu^T Q(z)ZY \\ 1 + \frac{1}{n}Y^T Z^T Q(z)\mu & -1 + \frac{1}{n}Y^T Z^T Q(z)\frac{1}{n}ZY\end{bmatrix}^{-1}\begin{bmatrix}\mu^T \\ \frac{1}{n}Y^T Z^T\end{bmatrix} Q(z)\left(\frac{1}{n}XY\right)$$

Como $X = \mu Y^T + Z$, lo anterior es idéntico a, $A - BCD$, donde

$$A = \mu^T Q(z)\mu + \frac{1}{n}\mu^T Q(z)ZY$$

$$B = \begin{bmatrix}\mu^T Q(z)\mu & \mu^T Q(z)\frac{1}{n}ZY\end{bmatrix}$$

$$C = \begin{bmatrix}\mu^T Q(z)\mu & 1 + \frac{1}{n}\mu^T Q(z)ZY \\ 1 + \frac{1}{n}Y^T Z^T Q(z)\mu & -1 + \frac{1}{n}Y^T Z^T Q(z)\frac{1}{n}ZY\end{bmatrix}^{-1}$$

$$D = \begin{bmatrix}\mu^T Q(z)\mu + \mu^T Q(z)\frac{1}{n}ZY \\ \frac{1}{n}Y^T Z^T Q(z)\mu + \frac{1}{n}Y^T Z^T Q(z)\frac{1}{n}ZY\end{bmatrix}$$

Posteriormente, nos enfocamos en la expresión de $\mu^T(\frac{1}{n}XX^T - zI_p)^{-1}\frac{1}{n}XY$. Podemos ver que los terminos que estan involucrados son unícamente $\mu^T Q(z)\mu$, $\frac{1}{n}\mu^T Q(z)ZY$ y $\frac{1}{n^2}Y^T Z^T Q(z)ZY$. Es por ello que, por la Proposición 3.1.2, se obtiene que $\mu^T(\frac{1}{n}XX^T - zI_p)^{-1}\frac{1}{n}XY$ converge con probabilidad 1, cuando $n \to \infty$, a

$$m(z)||\mu||^2 - \begin{bmatrix}||\mu||^2 m(z) & 0\end{bmatrix}\begin{bmatrix}||\mu||^2 m(z) & 1 \\ 1 & z\tilde{m}(z)\end{bmatrix}^{-1}\begin{bmatrix}m(z)||\mu||^2 \\ 1 + z\tilde{m}(z)\end{bmatrix}.$$

Ahora, sacamos la inversa de la matriz anterior, para llegar a

$$m(z)||\mu||^2 - \frac{1}{z||\mu||^2 m(z)\tilde{m}(z) - 1}\begin{bmatrix}||\mu||^2 m(z) & 0\end{bmatrix}\begin{bmatrix}z\tilde{m}(z) & -1 \\ -1 & ||\mu||^2 m(z)\end{bmatrix}\begin{bmatrix}m(z)||\mu||^2 \\ 1 + z\tilde{m}(z)\end{bmatrix}$$

y esto es,

$$= \frac{z||\mu||^4 m^2(z)\tilde{m}(z) - m(z)||\mu||^2 - z||\mu||^4 m^2(z)\tilde{m}(z) + m(z)||\mu||^2 + z||\mu||^2 m(z)\tilde{m}(z)}{z||\mu||^2 m(z)\tilde{m}(z) - 1}$$

lo cual es lo mismo que,

$$\frac{z\tilde{m}(z)||\mu||^2 m(z)}{z||\mu||^2 m(z)\tilde{m}(z) - 1}.$$

Reducimos la expresión anterior, el limite de $\mu^T \left(\frac{1}{n}XX^T - zI_p\right)^{-1} \frac{1}{n}XY$, de la siguiente manera: Por la Proposición 1.1.5, $(zm(z)+1)(cm(z)+1) = m(z)$, por lo que

$$\begin{aligned} & zcm^2(z) + zm(z) + cm(z) + 1 = m(z) \\ \Rightarrow\ & zcm^2(z) + zm(z) - m(z) + cm(z) + 1 = 0 \\ \Rightarrow\ & m(z)(zcm(z) - 1 + c) + zm(z) + 1 = 0. \end{aligned}$$

Por otro lado, por Proposición 1.1.7, se tiene que $\tilde{m}(z) = cm(z) - \frac{1}{z}(1-c)$, de donde, lo anterior implica que

$$zm(z)\tilde{m}(z) + zm(z) + 1 = 0$$

multiplicando por $||\mu||^2$,

$$\begin{aligned} & ||\mu||^2(zm(z)\tilde{m}(z) + zm(z) + 1) = 0 \\ \Rightarrow\ & ||\mu||^2 zm(z)\tilde{m}(z) + ||\mu||^2 zm(z) + ||\mu||^2 = 0 \\ \Rightarrow\ & ||\mu||^2 zm(z)\tilde{m}(z) = -||\mu||^2(zm(z)+1) \end{aligned}$$

sumando 0 ambos lado de la ecuación anterior,

$$||\mu||^2 zm(z)\tilde{m}(z) + ||\mu||^4 z^2 m^2(z)\tilde{m}(z) + ||\mu||^4 zm(z)\tilde{m}(z) = -||\mu||^2(zm(z)+1) + ||\mu||^4 z^2 m^2(z)\tilde{m}(z) + ||\mu||^4 zm(z)\tilde{m}(z)$$

de donde, sacando factor común

$$(||\mu||^2 zm(z)\tilde{m}(z))(1 + ||\mu||^2(zm(z)+1)) = (||\mu||^2(zm(z)+1))(||\mu||^2 zm(z)\tilde{m}(z) - 1)$$

lo cual implica,

$$\frac{z\tilde{m}(z)||\mu||^2 m(z)}{z||\mu||^2 m(z)\tilde{m}(z) - 1} = \frac{||\mu||^2(zm(z)+1)}{1 + ||\mu||^2(zm(z)+1)}.$$

Ahora, del hecho de que $zm(z) + 1 = \frac{m(z)}{cm(z)+1}$, (Proposición 1.1.5)

$$\begin{aligned} \frac{||\mu||^2(zm(z)+1)}{1 + ||\mu||^2(zm(z)+1)} &= \frac{\frac{||\mu||^2 m(z)}{(cm(z)+1)}}{1 + \frac{||\mu||^2 m(z)}{cm(z)+1}} \\ &= \frac{\frac{||\mu||^2 m(z)}{(cm(z)+1)}}{\frac{cm(z)+1+||\mu||^2 m(z)}{cm(z)+1}} \\ &= \frac{||\mu||^2 m(z)}{m(z)(c + ||\mu||^2) + 1}. \end{aligned}$$

Por lo que de esta manera se tiene el resultado. □

El Lema 3.1.5 concede la convergencia de uno de los sumandos encontrados en la formula de $\mu^T w(t)$ en el Lema 3.1.2. Para el otro sumando tenemos el Lema 3.1.6.

Lema 3.1.6. *Cuando $n \to \infty$, con probabilidad 1,*

$$\mu^T \left(\frac{1}{n} XX^T - zI_p \right)^{-1} w_0 \to 0.$$

Demostración. Por el Lema 3.1.4, $\mu^T \left(\frac{1}{n} XX^T - zI_p \right)^{-1} w_0$ es igual a

$$\mu^T Q(z) w_0 - \mu^T Q(z) \begin{bmatrix} \mu & \frac{1}{n} ZY \end{bmatrix} \begin{bmatrix} \mu^T Q(z) \mu & 1 + \frac{1}{n} \mu^T Q(z) ZY \\ 1 + \frac{1}{n} Y^T Z^T Q(z) \mu & -1 + \frac{1}{n} Y^T Z^T Q(z) \frac{1}{n} ZY \end{bmatrix}^{-1} \begin{bmatrix} \mu^T \\ \frac{1}{n} Y^T Z^T \end{bmatrix} Q(z) w_0$$

lo cual se puede reescribir como,

$$\mu^T Q(z) w_0 - \begin{bmatrix} \mu^T Q(z) \mu & \mu^T Q(z) \frac{1}{n} ZY \end{bmatrix} \begin{bmatrix} \mu^T Q(z) \mu & 1 + \frac{1}{n} \mu^T Q(z) ZY \\ 1 + \frac{1}{n} Y^T Z^T Q(z) \mu & -1 + \frac{1}{n} Y^T Z^T Q(z) \frac{1}{n} ZY \end{bmatrix}^{-1} \begin{bmatrix} \mu^T Q(z) w_0 \\ \frac{1}{n} Y^T Z^T Q(z) w_0 \end{bmatrix}$$

Nuevamente, al igual que en el Lema 3.1.5, se obtuvieron unícamente terminos de los cuales sabemos su comportamiento asintótico. Exactamente, haciendo uso de la Proposición 3.1.2 se sigue la convergencia con probabilidad 1, cuando $n \to \infty$, de $\mu^T \left(\frac{1}{n} XX^T - zI_p \right)^{-1} w_0$ a

$$0 - \begin{bmatrix} ||\mu||^2 m(z) & 0 \end{bmatrix} \begin{bmatrix} ||\mu||^2 m(z) & 1 \\ 1 & z\tilde{m}(z) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0.$$

□

Hasta este momento, por medio del Lema 3.1.5 y 3.1.6 se tiene la convergencia de $\mu^T w(t)$. A continuación se establecen lemas que nos dan la convergencia de $w(t)^T w(t)$ y de esta manera poder pasar a dar explicitamente la demostración del Teorema 3.1.1.

Lema 3.1.7. *Cuando $n \to \infty$, con probabilidad 1, $w_0^T \left(\frac{1}{n} XX^T - zI_p \right)^{-1} w_0$ converge a $\sigma^2 m(z)$.*

Demostración. Tenemos por, Lema 3.1.4, que

$$w_0^T Q(z) w_0 - w_0^T Q(z) \begin{bmatrix} \mu & \frac{1}{n} ZY \end{bmatrix} \begin{bmatrix} \mu^T Q(z) \mu & 1 + \frac{1}{n} \mu^T Q(z) ZY \\ 1 + \frac{1}{n} Y^T Z^T Q(z) \mu & -1 + \frac{1}{n} Y^T Z^T Q(z) \frac{1}{n} ZY \end{bmatrix}^{-1} \begin{bmatrix} \mu^T \\ \frac{1}{n} Y^T Z^T \end{bmatrix} Q(z) w_0$$

y esto es,

$$w_0^T Q(z) w_0 - \begin{bmatrix} w_0^T Q(z) \mu & w_0^T Q(z) \frac{1}{n} ZY \end{bmatrix} \begin{bmatrix} \mu^T Q(z) \mu & 1 + \frac{1}{n} \mu^T Q(z) ZY \\ 1 + \frac{1}{n} Y^T Z^T Q(z) \mu & -1 + \frac{1}{n} Y^T Z^T Q(z) \frac{1}{n} ZY \end{bmatrix}^{-1} \begin{bmatrix} \mu^T Q(z) w_0 \\ \frac{1}{n} Y^T Z^T Q(z) w_0 \end{bmatrix}$$

De igual forma a como lo hemos estado realizando, por la Proposición 3.1.2, resulta que $w_0^T \left(\frac{1}{n} XX^T - zI_p \right)^{-1} w_0$ converge a

$$\sigma^2 m(z) - \begin{bmatrix} 0 & 0 \end{bmatrix} \begin{bmatrix} m(z)||\mu|| & 1 \\ 1 & z\tilde{m}(z) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \sigma^2 m(z)$$

con probabilidad 1, cuando $n \to \infty$. □

Lema 3.1.8. *Se tiene la convergencia*

$$\frac{1}{n}Y^TX^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY\to 1-\frac{1}{(||\mu||^2+c)m(z)+1}$$

con probabilidad 1, cuando $n\to\infty$.

Demostración. Reiteradamente por Lema 3.1.4, $\frac{1}{n}Y^TX^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$ se puede escribir como la siguiente diferencia,

$$\frac{1}{n}Y^TX^TQ(z)\frac{1}{n}XY$$

$$-\frac{1}{n}Y^TX^TQ(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix}\mu^TQ(z)\mu & 1+\frac{1}{n}\mu^TQ(z)ZY\\ 1+\frac{1}{n}Y^TZ^TQ(z)\mu & -1+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}^{-1}\begin{bmatrix}\mu^T\\ \frac{1}{n}Y^TZ^T\end{bmatrix}Q(z)\frac{1}{n}XY.$$

Ahora, puesto que $X=\mu Y^T+Z$, lo anterior puede ser reescrito de la siguiente manera $A-BCD$, donde

$$A=\mu^TQ(z)\mu+\mu^TQ(z)\frac{1}{n}ZY+\frac{1}{n}Y^TZ^TQ(z)\mu+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY$$

$$B=\begin{bmatrix}\mu^TQ(z)\mu+\frac{1}{n}Y^TZ^TQ(z)\mu & \frac{1}{n}Y^TZ^TQ(z)\mu+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}$$

$$C=\begin{bmatrix}\mu^TQ(z)\mu & 1+\frac{1}{n}\mu^TQ(z)ZY\\ 1+\frac{1}{n}Y^TZ^TQ(z)\mu & -1+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}^{-1}$$

$$D=\begin{bmatrix}\mu^TQ(z)\mu+\mu^TQ(z)\frac{1}{n}ZY\\ \frac{1}{n}Y^TZ^TQ(z)\mu+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}.$$

De aquí por la Proposición 3.1.2 $\frac{1}{n}Y^TX^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$ converge con probabilidad 1, cuando $n\to\infty$ a

$$m(z)||\mu||^2+1+z\tilde{m}(z)-\begin{bmatrix}m(z)||\mu||^2 & 1+z\tilde{m}(z)\end{bmatrix}\begin{bmatrix}m(z)||\mu||^2 & 1\\ 1 & z\tilde{m}(z)\end{bmatrix}^{-1}\begin{bmatrix}m(z)||\mu||^2\\ 1+\tilde{m}(z)\end{bmatrix}$$

Luego, desarrolando la expresión anterios se obtiene

$$m(z)||\mu||^2+1+z\tilde{m}(z)-\begin{bmatrix}m(z)||\mu||^2 & 1+z\tilde{m}(z)\end{bmatrix}\frac{1}{z\tilde{m}(z)m(z)||\mu||^2-1}\begin{bmatrix}z\tilde{m}(z) & -1\\ -1 & m(z)||\mu||^2\end{bmatrix}\begin{bmatrix}m(z)||\mu||^2\\ 1+\tilde{m}(z)\end{bmatrix}$$

despues de efectuar las multiplicaciones de las matrices involucradas en la expresión anterior, resulta

$$m(z)||\mu||^2+1+z\tilde{m}(z)-\frac{z\tilde{m}(z)m^2(z)||\mu||^4-||\mu||^2m(z)-z\tilde{m}(z)m(z)||\mu||^2+z^2\tilde{m}^2(z)m(z)||\mu||^2+z\tilde{m}(z)m(z)||\mu||^2}{z\tilde{m}(z)m(z)||\mu||^2-1}$$

y esto es identico a,

$$m(z)||\mu||^2+1+z\tilde{m}(z)-\frac{(z\tilde{m}(z)m(z)||\mu||^2-1)(||\mu||^2m(z))+(z\tilde{m}(z)m(z)||\mu||^2-1)(z\tilde{m}(z))+z\tilde{m}(z)}{z\tilde{m}(z)m(z)||\mu||^2-1}$$

lo cual, finalmente, es equivalente a,

$$1-\frac{z\bar{m}(z)}{z\tilde{m}(z)m(z)||\mu||^2-1}=1-\frac{z\bar{m}(z)||\mu||^2m(z)}{z\tilde{m}(z)m(z)||\mu||^2-1}\frac{1}{||\mu||^2m(z)}$$

por lo visto en el Lema 3.1.7,esto es

$$1-\frac{||\mu||^2m(z)}{(||\mu||^2+c)m(z)+1}\frac{1}{||\mu||^2m(z)}$$

y esto es exactamente lo deseado,

$$1-\frac{1}{(||\mu||^2+c)m(z)+1}.$$

□

Lema 3.1.9. *Cuando $n\to\infty$, $w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$ converge con probabilidad 1 a cero.*

Demostración. Haciendo uso del Lema 3.1.4, reescribimos $w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$ de la siguiente forma

$$w_0^TQ(z)\frac{1}{n}XY-w_0^TQ(z)\begin{bmatrix}\mu & \frac{1}{n}ZY\end{bmatrix}\begin{bmatrix}\mu^TQ(z)\mu & 1+\frac{1}{n}\mu^TQ(z)ZY\\ 1+\frac{1}{n}Y^TZ^TQ(z)\mu & -1+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}^{-1}\begin{bmatrix}\mu^T\\ \frac{1}{n}Y^TZ^T\end{bmatrix}Q(z)\frac{1}{n}XY.$$

Ahora, desarrollamos la expresión anterior, como lo hemos estado haciendo en los lemas anteriores, obtenemos que $w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$ es igual a $A-BCD$, donde

$$A=w_0^TQ(z)\mu+w_0^TQ(z)\frac{1}{n}ZY$$

$$B=\begin{bmatrix}w_0^TQ(z)\mu & w_0^TQ(z)\frac{1}{n}ZY\end{bmatrix}$$

$$C=\begin{bmatrix}\mu^TQ(z)\mu & 1+\frac{1}{n}\mu^TQ(z)ZY\\ 1+\frac{1}{n}Y^TZ^TQ(z)\mu & -1+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}^{-1}$$

$$D=\begin{bmatrix}\mu^TQ(z)\mu+\mu^TQ(z)\frac{1}{n}ZY\\ \frac{1}{n}Y^TZ^TQ(z)\mu+\frac{1}{n}Y^TZ^TQ(z)\frac{1}{n}ZY\end{bmatrix}$$

resultando una formula, para $w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$, en la cual todos los terminos tienen un comportamiento asintótico en probabilidad conocido. Así que, $w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XY$ converge a

$$0-\begin{bmatrix}0 & 0\end{bmatrix}\begin{bmatrix}m(z)||\mu||^2 & 1\\ 1 & z\tilde{m}(z)\end{bmatrix}^{-1}\begin{bmatrix}m(z)||\mu||^2\\ 1+z\tilde{m}(z)\end{bmatrix}=0$$

con probabilidad 1, cuando $n\to\infty$. □

A partir de este punto, se cuenta con todas las herramientas para proceder a escribir la demostración del resultado principal de esta sección.

Teorema 3.1.10. *Sea $(x, y) \in D \times \mathcal{Y}$, no dato de entrenamiento. Cuando $n \to \infty$; con probabilidad* 1

$$P\left(w(t)^T x > 0 \mid y = -1\right) - Q\left(E/\sqrt{V}\right) \to 0$$

$$P\left(w(t)^T x < 0 \mid y = 1\right) - Q\left(E/\sqrt{V}\right) \to 0$$

donde:

$$E = \frac{-1}{2i\pi} \oint_\gamma \frac{1 - f_t(z)}{z} \frac{||\mu||^2 m(z)}{(||\mu||^2 + c)m(z) + 1} dz$$

$$V = \frac{1}{2i\pi} \oint_\gamma \left(\frac{\frac{1}{z^2}(1 - f_t(z))^2}{(||\mu||^2 + c)m(z) + 1} - \sigma^2 (f_t(z))^2 m(z) \right) dz$$

con $f_t(z) = exp(-\alpha t z)$, $m(z) = \frac{1-c-z}{2cz} + \frac{\sqrt{(1-c-z)^2 - 4cz}}{2cz}$ la transformada de stieltjes de la distribución marchenko-pastur (Proposición 1.1.4), $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp(-\frac{u^2}{2})\, \mathrm{d}u$ la función conocida como Q-function y γ un camino cerrado orientado positivamente que contiene todos los valores propios de $\frac{1}{n} X X^T$ y al origen.

Demostración. **(Demostración Teorema 3.1.1)**
Por la Proposición 3.1.1,

$$P\left(w(t)^T x > 0 \mid y = -1\right) = P\left(w(t)^T x < 0 \mid y = 1\right) = Q\left(\frac{\mu^T w(t)}{||w(t)||}\right)$$

de manera que basta encontrar el limite cuando, $n \to \infty$, de $Q\left(\frac{\mu^T w(t)}{||w(t)||}\right)$ casi seguramente. Por la continuidad de la $Q\text{-}function$ y la función raiz cuadrada, conociendo el comportamiento asintótico casi seguramente de $\mu^T w(t)$ y $w(t)^T w(t)$, acabamos. Para ello, mediante el Lema 3.1.2 y 3.1.3 encontramos que

$$\mu^T w(t) = -\frac{1}{2i\pi} \oint_\gamma f_t(z) \mu^T \left(\frac{1}{n} X X^T - z I_p\right)^{-1} w_0 dz$$

$$-\frac{1}{2i\pi} \oint_\gamma \frac{(1 - f_t(z))}{z} \mu^T \left(\frac{1}{n} X X^T - z I_p\right)^{-1} \frac{1}{n} X Y dz$$

mientras que

$$w(t)^T w(t) = -\frac{1}{2i\pi} \oint_\gamma f_t^2(z) w_0^T \left(\frac{1}{n} X X^T - z I_p\right)^{-1} w_0 dz$$

$$-\frac{1}{i\pi} \oint_\gamma \frac{f_t(z)(1 - f_t(z))}{z} w_0^T \left(\frac{1}{n} X X^T - z I_p\right)^{-1} \frac{1}{n} X Y dz$$

$$-\frac{1}{2i\pi}\oint_\gamma \frac{(1-f_t(z))^2}{z^2}\frac{1}{n}Y^TX^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XYdz.$$

Observemos ahora que haciendo uso del Teorema de Convergencia Dominada de Teoria de la Medida, teniendo el límite casi seguro de las sucesiones dentro de las integrales en $\mu^T w(t)$ y $w(t)^t w(t)$ respectivamente, resulta el comportamiento asintótico de $\mu^T w(t)$ y de $w(t)^t w(t)$ casi seguro. De modo que, por Lema 3.1.5 y 3.1.6

$$\mu^T w(t) = -\frac{1}{2i\pi}\oint_\gamma f_t(z)\mu^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}w_0dz - \frac{1}{2i\pi}\oint_\gamma \frac{(1-f_t(z))}{z}\mu^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XYdz$$

$$\to -\frac{1}{2i\pi}\oint_\gamma f_t(z)0dz - \frac{1}{2i\pi}\oint_\gamma \frac{(1-f_t(z))}{z}\frac{||\mu||^2 m(z)}{m(z)(c+||\mu||^2)+1}dz$$

$$= -\frac{1}{2i\pi}\oint_\gamma \frac{(1-f_t(z))}{z}\frac{||\mu||^2 m(z)}{m(z)(c+||\mu||^2)+1}dz = E.$$

Por otro lado, es consecuencia de los Lemas 3.1.7, 3.1.8 y 3.1.9 que

$$w(t)^T w(t) = -\frac{1}{2i\pi}\oint_\gamma f_t^2(z)w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}w_0dz$$

$$-\frac{1}{i\pi}\oint_\gamma \frac{f_t(z)(1-f_t(z))}{z}w_0^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XYdz$$

$$-\frac{1}{2i\pi}\oint_\gamma \frac{(1-f_t(z))^2}{z^2}\frac{1}{n}Y^TX^T\left(\frac{1}{n}XX^T-zI_p\right)^{-1}\frac{1}{n}XYdz$$

$$\to -\frac{1}{2i\pi}\oint_\gamma f_t^2(z)\sigma^2 m(z)dz - \frac{1}{i\pi}\oint_\gamma \frac{f_t(z)(1-f_t(z))}{z}0dz$$

$$-\frac{1}{2i\pi}\oint_\gamma \frac{(1-f_t(z))^2}{z^2}\left(1-\frac{1}{(||\mu||^2+c)m(z)+1}\right)dz$$

esto ultimo es identico a,

$$\frac{1}{2i\pi}\oint_\gamma\left(\frac{\frac{1}{z^2}(1-f_t(z))^2}{(||\mu||^2+c)m(z)+1}-\sigma^2(f_t(z))^2m(z)-\frac{1}{z^2}(1-f_t(z))^2\right)dz = V.$$

Cabe mencionar que las convergencias anteriores son cuando $n\to\infty$ con prababilidad 1. De aquí, observando que $||w(t)|| = \sqrt{w(t)^T w(t)} \to \sqrt{V}$, concluimos que

$$Q\left(\frac{\mu^T w(t)}{||w(t)||}\right) \to Q\left(\frac{E}{\sqrt{V}}\right)$$

mediante lo cual se completa la demostración del Teorema 3.1.1. □

3.2. Rendimiento de Generalización Vía Probabilidad Libre

Se exponen algunas replicas que se lograron en esta tesis, de pasos en la prueba del resultado principal del trabajo de Liao y Couillet [18], el Teorema 3.1.1, que fueron obtenidos al utilizar la teoría de probabilidad libre, concretamente su extensión infinitesimal, en lugar de equivalentes deterministicos como lo hacen Liao y Couillet [18]. Recordemos que en la prueba del resultado principal (Sección 3.1), de manera especifica, en la demostración de el Lema 3.1.2, el uso de equivalentes deterministicos fue fundamental para la obtención del resultado principal. De hecho, resulto un tanto acomodaticio al problema, como lo muestra el punto **1** de el Lema 3.1.2, e informal, al usar $Q(z) = m(z)I_p$ pues en realidad son distintos, por ejemplo no tienen los mismos valores propios. De este modo, obtenemos cierta rigurosidad en la demostración del Teorema 3.1.1.

Acontinuación se demuestran los puntos **1** y **6** de el Lema 3.1.2 utilizando la teoría de distribución infinitesimal.

Proposición 3.2.1. *Consideremos las matrices $\mu\mu^T$ y $Q(z)$, las cuales son matrices de tamaño $p \times p$. Entonces, el ensamble $\left\{Q(z)\mu\mu^T\right\}$ tiene distribución infinitesimal.*

Demostración. Sabemos de la Definición 1.4.3 que basta demostrar
i). $\phi(x^k) = \lim_{p\to\infty} \mathbb{E}(\frac{1}{p}Tr((Q(z)\mu\mu^T)^k))$ y
ii). $\phi'(x^k) = \lim_{p\to\infty} p(\mathbb{E}(\frac{1}{p}Tr((Q(z)\mu\mu^T)^k)) - \phi(x^k))\,, k \in \mathbb{N}$.
Procederemos a la prueba de 1.
Para ello, sea $k \in \mathbb{N}$. Primero observemos que,

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{p}Tr((\mu\mu^T)^k)\right) &= \mathbb{E}\left(\frac{1}{p}Tr(\mu\mu^T ... \mu\mu^T)\right) && \text{(k-veces)} \\
&= \mathbb{E}\left(\frac{1}{p}Tr(\mu^T\mu ... \mu^T\mu)\right) && \text{(k-veces)} \\
&= \frac{1}{p}||\mu||^{2k}.
\end{aligned}$$

Por otro lado, puesto que $||\mu||^2 = O(1)$ existe una constante C tal que $0 \leq ||\mu||^2 \leq C$ para todo p. De modo que,

$$0 \leq \lim_{p\to\infty} \frac{1}{p}||\mu||^{2k} \leq \lim_{p\to\infty} \frac{C^k}{p} = 0$$

y por lo tanto $\lim_{p\to\infty} \mathbb{E}\left(\frac{1}{p}Tr((\mu\mu^T)^k)\right) = 0$, es decir la matriz $\mu\mu^T$ tiene distribución límite igual 0. De aquí, como Z es una matriz aleatoria gaussiana, por el Teorema 1.3.1 del Capítulo 1, se sigue que las matrices Z y $\mu\mu^T$ son asíntoticamente libres. Más aun, por definición de libertad, el algebra generada por la matriz Z es asintóticamente libre del algebra generada por $\mu\mu^T$. Así que, debido a la forma que tiene el resolvente de la matriz

$\frac{1}{n}ZZ^T$, se obtiene que $\mu\mu^T$ y $Q(z)$ son asíntoticamente libres. Ahora, por el Teorema 1.3.4, la fórmula en terminos de cumulantes libres de productos de matrices aleatorias libres,

$$\begin{aligned}\lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr\left(\left(Q(z)\mu\mu^T\right)^k\right)\right) &= \lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr\left(Q(z)\mu\mu^T...Q(z)\mu\mu^T\right)\right) \qquad \text{(k-veces)}\\ &= \sum_{\pi\in NC(n)}\kappa_\pi(Q(z),...,Q(z))\cdot\left(\lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr\right)_{Kr(\pi)}(\mu\mu^T,...,\mu\mu^T)\right)\\ &= 0\end{aligned}$$

donde la última igualdad se da porque los cumulantes libres de $\mu\mu^T$ son todos igual a cero, ya que la distribución límite de $\mu\mu^T$ es 0 (se ve de la definición de cumulantes libres, 1.3.9). Esto muestra que $\phi(x^k)=0$.
Para examinar $ii)$, notemos que por lo demostrado en 1, basta encontrar $\lim_{p\to\infty}E(Tr((Q(z)\mu\mu^T)^k))$. Para ello, puesto que μ es un vector determinista con $||\mu||=O(1)$, sin pérdida de generalidad se supone que μ es un vector con únicamente la primera entrada no cero.
Denotando $A_{i,j}$ la entrada (i,j) de la matriz A, tenemos que

$$\begin{aligned}Tr((Q(z)\mu\mu^T)^k) &= \sum_{i_1,...,i_k,j_1,...,j_k=1}^{p}Q(z)_{i_1,j_1}\mu\mu^T_{j_1,i_2}...Q(z)_{i_k,j_k}\mu\mu^T_{j_k,i_1} &(3.12)\\ &= (Q(z)_{1,1})^k(\mu\mu^T_{1,1})^k &(3.13)\\ &= (Q(z)_{1,1})^k||\mu||^{2k} &(3.14)\end{aligned}$$

donde la ecuación (3.13) se da ya que la matriz $\mu\mu^T$ tiene todas sus entradas igual a cero excepto la primera. Ahora, por la Proposición 1.1.7

$$\lim_{p\to\infty}Q(z)_{1,1}=m(z)$$

de donde, usando la continuidad de la función x^k,

$$\lim_{p\to\infty}Q(z)^k_{1,1}=m(z)^k.$$

Además, por la Proposición 1.1.3, sabemos que el resolvente $Q(z)$ tiene norma de operador acotado, lo cual implica $Q(z)_{1,1}$ es acotado. De modo que, por el Teorema de Convergencia Dominada,

$$\lim_{p\to\infty}\mathbb{E}(Q(z)^k_{1,1})=m(z)^k.$$

Así que,

$$\lim_{p\to\infty}\mathbb{E}(Tr((Q(z)\mu\mu^T)^k))=\lim_{p\to\infty}\mathbb{E}((Q(z)_{1,1})^k)||\mu||^{2k}=m(z)^k||\mu||^{2k}.$$

Concluimos de esta manera $ii)$. □

Haciendo uso de la Proposición 3.2.1, ya que $\mathbb{E}(\mu^T Q(z)\mu) = \mathbb{E}(Tr(\mu^T Q(z)\mu)) = \mathbb{E}(Tr(Q(z)\mu\mu^T))$, resulta que $\mu^T Q(z)\mu$ converge en distribución infinitesimal con,

$$\lim_{p\to\infty} \mathbb{E}(\mu^T Q(z)\mu) = \lim_{p\to\infty} \mathbb{E}(Tr(Q(z)\mu\mu^T)) = m(z)||\mu||^2.$$

obteniendo de esta manera una prueba de el punto **1.** de el Lema 3.1.2 en la que se evita el uso de equivalentes deterministicos de la manera en que lo realizo Liao y Couillet en su trabajo.

Proposición 3.2.2. *Consideremos las matrices $w_0 w_0^T$ y $Q(z)$, las cuales son matrices de tamaño $p \times p$. Entonces, el ensamble $\left\{Q(z)w_0 w_0^T\right\}$ tiene distribución infinitesimal.*

Demostración. Sabemos de la Definición 1.4.3 que se debe demostrar
i). $\phi(x^k) = \lim_{p\to\infty} \mathbb{E}(\frac{1}{p}Tr((Q(z)w_0 w_0^T)^k))$ y
ii). $\phi'(x^k) = \lim_{p\to\infty} p(\mathbb{E}(\frac{1}{p}Tr((Q(z)w_0 w_0^T)^k)) - \phi(x^k))\ , k \in \mathbb{N}$.
Sea $k \in \mathbb{N}$. Iniciamos con la prueba de 1.
Tenemos que,

$$\begin{aligned}
\frac{1}{p}Tr\left(\left(w_0 w_0^T\right)^k\right) &= \frac{1}{p}Tr\left(\left(w_0 w_0^T ....w_0 w_0^T\right)\right) && \text{(k-veces)}\\
&= \frac{1}{p}Tr\left(\left(w_0^T w_0 ....w_0^T w_0\right)\right) && \text{(k-veces)}\\
&= \frac{1}{p}||w_0||^{2k}
\end{aligned}$$

y puesto que se demostró en la Sección 3.1 que $||w_0||^2 \to \sigma^2$ cuando $p \to \infty$, se sigue que $||w_0||^2 = O(1)$ y por lo tanto $||w_0||^2 \leq C$ para todo p, con C constante. De modo que,

$$0 \leq \lim_{p\to\infty} \frac{1}{p}||w_0||^2 \leq \lim_{p\to\infty} \frac{1}{p}C = 0$$

lo cual muestra que $\lim_{p\to\infty} \frac{1}{p}Tr\left(\left(w_0 w_0^T\right)^k\right) = 0$ y en consecuencia $w_0 w_0^T$ converge en distribución al 0. De aquí, como la matriz Z es aleatoria gaussiana e independiente a w_0, por la Proposición 1.3.2 del Capítulo 1, se sigue que Z y $w_0 w_0^T$ son asíntoticamente libres. Por lo que, por la forma que tiene la matriz $Q(z)$, $w_0 w_0^T$ y $Q(z)$ son asíntoticamente libres. Nuevamente, haciendo uso del Teorema 1.3.4 como en la Proposición 3.2.1, se muestra que $\phi(x^k) = 0$.
Para la demostración de *ii*) observemos que únicamente se tiene que analizar la existencia del límite $\mathbb{E}(Tr((Q(z)w_0 w_0^T)^k))$. Así que necesitamos considerar los siguientes casos.
i) $k = 1$ Entonces,

$$E(Tr(Q(z)w_0 w_0^T)) = \mathbb{E}(Tr(Q(z)\pi_{1,i}w_0(\pi_{1,i}w_0)^T))$$

con $\pi_{1,i}$ matriz de permutación $1 \leftrightarrow i$. De modo que,

$$E(Tr(Q(z)w_0 w_0^T)) = \mathbb{E}\left(\frac{1}{p}Tr\left(Q(z)\left(\sum_{i=1}^{k}(\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)\right)\right)$$

Ahora, sea $m \in \mathbb{N}$. Para $i \in \{1, .., p\}$, la matriz $(\pi_{1,i}w_0)(\pi_{1,i}w_0)^T$ es matriz semi-positiva definida y por lo tanto $\left((\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)^m$ es semi-positiva definida. De modo que sus valores propios son todos no negativos, y en consecuencia

$$0 \leq \mathbb{E}\left(\frac{1}{p}Tr\left(\left((\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)^m\right)\right)$$

Por otro lado,

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{p}Tr\left(\left((\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)^m\right)\right) &= \mathbb{E}\left(\frac{1}{p}Tr\left(\left(\pi_{1,i}(w_0w_0^T)\pi_{1,i}^T\right)^m\right)\right) && (3.15)\\
&\leq \mathbb{E}\left(\frac{1}{p}Tr\left(\pi_{1,i}(w_0w_0^T)^m\pi_{1,i}^T\right)\right) && (3.16)\\
&= \frac{||w_0||^{2m}}{p} && (3.17)
\end{aligned}$$

donde la ecuación (3.17) es debido a la desigualdad de Jensen para la traza (Apéndice B), ya que la función x^m es continua, convexa en $[0, \infty)$ y además el espectro de $w_0w_0^T$ (el cual es $\left\{||w_0||^2, 0\right\}$) esta contenido en este intervalo. De lo anterior, puesto que $||w_0||^{2m} \to \sigma^{2m}$ cuando $p \to \infty$, se sigue que $\lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr\left(\left((\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)^m\right)\right) = 0$. Luego, por el Teorema 1.3.2, $Q(z)$ y $(\pi_{1,i}w_0)(\pi_{1,i}w_0)^T$ son asintóticamente libres. Resulta entonces que,

$$\begin{aligned}
\lim_{p\to\infty}\mathbb{E}(Tr(Q(z)w_0w_0^T)) &= \lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr\left(Q(z)\left(\sum_{i=1}^{p}(\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)\right)\right)\\
&= \lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr(Q(z))\right)\lim_{p\to\infty}\mathbb{E}\left(\frac{1}{p}Tr\left(\sum_{i=1}^{p}(\pi_{1,i}w_0)(\pi_{1,i}w_0)^T\right)\right)\\
&= m(z)\sigma^2.
\end{aligned}$$

ii) $k > 1$ Se conjetura su validez. □

Observemos que, $\mathbb{E}(w_0^TQ(z)w_0) = \mathbb{E}(Tr(w_0^TQ(z)w_0)) = \mathbb{E}(Tr(Q(z)w_0w_0^T))$, por lo que la Proposición 3.2.2 nos dice que $w_0^TQ(z)w_0$ converge en distribución infinitesimal con,

$$\lim_{p\to\infty}\mathbb{E}(w_0^TQ(z)w_0) = \lim_{p\to\infty}\mathbb{E}(Tr(Q(z)w_0w_0^T)) = m(z)\sigma^2.$$

De esta manera, el punto **6.** de la Proposición 3.1.2 tendría una demostrración más rigurosa evitando el uso de equivalentes deterministicos, y por lo tanto se da un enfoque distinto a como lo realizo Liao y Couillet en su trabajo.

3.3. Implementación Computacional

En esta última sección ilustramos numéricamente el resultado principal en esta tesis. Liao y Couillet [18] no mencionan con exactitud el proceso de la implementación computacional

que ellos realizan. De manera que las imágenes que se obtuvieron en esta tesis no son exactamente las mismas que ellos presentan en [18], sin embargo, podemos observar la misma fenomenología, una mejora de la primer imagen a la ultima imagen de esta sección, corroborando el análisis realizado por Liao y Couillet.

Primero, recordemos que a partir del Teorema 3.1.1 podemos controlar el rendimiento de generalización de la red neuronal artificial de una capa mediante la inicialización de los parámetros c, σ, α y μ. La Figura 3.1 presenta el rendimiento de generalización y aprendizaje de la red neuronal artificial, cuando ésta utiliza los parámetros ($n = 256, p = 512, \mu = [2, 0_{p-1}], \sigma^2 = 0{,}1, \alpha = 0{,}01$). Estos parámetros fueron escogidos arbitrariamente, es decir, no tienen nada en particular.

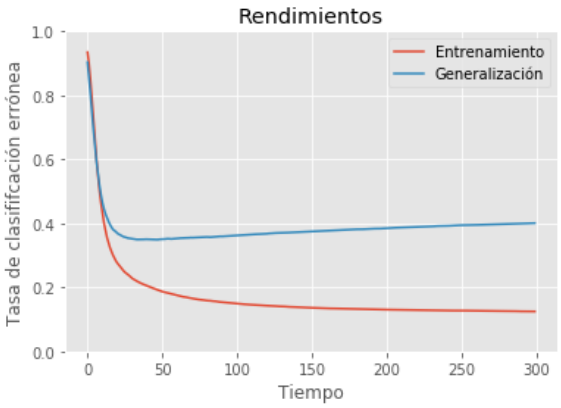

Figura 3.1: Rendimiento de la red neuronal

Se puede observar de la Figura 3.1 que el rendimiento de generalización es malo. Esto se da ya que la curva del rendimiento de generalización desciende rapidamente, sin embargo, en un punto entre los tiempos 0 y 50 comienza a crecer de forma brusca.

Ahora, la Figura 3.2 presenta el rendimiento de generalización y aprendizaje de la red neuronal artificial, cuando ésta utiliza datos reales, la base de datos MNIST (numeros 1 y 7) para la clasificación, con los parámetros ($n = 784, p = 784, \sigma^2 = 0{,}1, \alpha = 0{,}01$). Couillet y Liao a partir del Teorema 3.1.1 observaron ciertos indicios matemáticos que no explican a profundidad, con los cuales realizaron una cantidad grande de experimentos computacionales, hasta alcanzar los parametros de la Figura 3.2, que son los que proporcionan el mejor rendimiento de generalización con una red neuronal artificial muy bien entrenada.

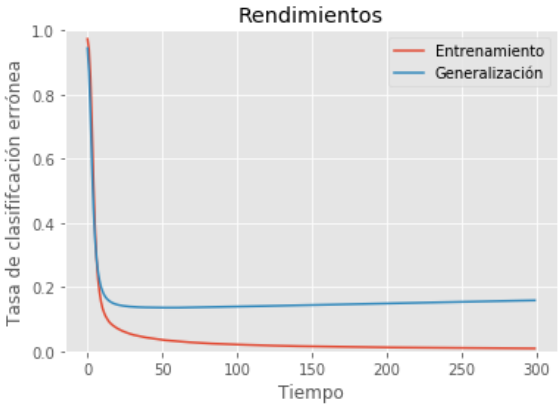

Figura 3.2: Rendimiento de la red neuronal

De hecho, en la Figura 3.2, se observa un rendimiento de generalización suave, lo cual es lo deseado. Mejorando drasticamente la Figura 3.1.

Se concluye entonces que a partir del Teorema 3.1.1, el resultado principal en [18], se puede controlar el rendimiento de generalización de la red neuronal artificial, mejorando o empeorando, con la elección de los parámetros de entrada a la red neuronal artificial que involucra dicho teorema. Más aun, la Figura 3.2 nos dice que haciendo uso de una parada anticipada en el entrenamiento de la red neuronal entre los tiempos 150 y 200 logramos solucionar una de las principales causas al obtener malos resultados a la hora de la generalización en Machine Learning, el sobre entrenamiento. Se observa que deteniendo el entrenamiento en un tiempo en el intervalo (150,200) logramos tener un rendimiento de generalización estable con un rendimiento de entrenamiento casi cero, lo cual es lo deseado.

Se proporciona entonces una validación computacional del Teorema 3.1.1.

3.3.1. Notas Adicionales

El Apéndice C contiene el código de Python con el cual se logran las imagenes del rendimiento de generalización y aprendizaje de la red neuronal artificial de una capa que se utiliza en este trabajo.

Apéndices

Apéndice A

Probabilidad Clásica

Este apéndice se incluye para facilitar la lectura de esta tesis. Daremos las definiciones y resultados relacionados con probabilidad clasica, ya que los temas abordados en esta tesis pueden interesar a lectores que no estén familiarizados con probabilidad. El contenido puede ser encontrados en gran variedad de libros, como por ejemplo [25].

El objetivo de la Teoría de Probabilidad es desarrollar y estudiar modelos matemáticos para experimentos cuyos resultados no pueden predecirse con exactitud, es decir aleatorios. Se puede decir que no fue hasta el siglo XX cuando esta teoría alcanzó un desarrollo axiomático. En 1933, A. N. Kolmogorov propone una axiomatización usando las ideas de la Teoría de Medida, desarrollada a principios del siglo XX por H. Lebesgue.

Empecemos recordando lo que se conoce como un espacio de probabilidad clásica.

Definición A.0.1. Un espacio de probabilidad es una terna $(\Omega, \mathcal{F}, \mathbb{P})$ donde Ω es un conjunto no vacio, $\mathcal{F}$ es una σ-álgebra de subconjuntos de Ω y $\mathbb{P}$ es una medida de probabilidad. Es decir, $\mathcal{F}$ cumple las propiedades:
i). $\Omega \in \mathcal{F}$.
ii). $A_1, A_2 ... \in \mathcal{F}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
iii). $A \in \mathcal{F}$, entonces $A^c \in \mathcal{F}$.

Mientras que $\mathbb{P} : \mathcal{F} \longrightarrow [0, 1]$ es tal que
$A1$. $\mathbb{P}(\Omega) = 1$.
$A2$. $\mathbb{P}$ es σ-aditiva: $A_1, A_2 ... \in \mathcal{F}$ y $A_i \cap A_j = \emptyset$ siempre que $i \neq j$, entonces

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Denotaremos por $\mathcal{B}(\mathbb{R})$ a la colección de todos los conjuntos de Borel en $\mathbb{R}$. Notemos que esta collecion es la σ-álgebra generada por los intervalos abiertos en $\mathbb{R}$. Decimos que $f : \mathbb{R} \longrightarrow \mathbb{R}$ es medible, si es $\mathcal{B}(\mathbb{R})-$medible.

Definición A.0.2. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Una función $X : \Omega \longrightarrow \mathbb{R}$ es una variable aleatoria, si es $\mathcal{F}$-medible en $\mathbb{R}$, esto es, $\{w : X(w) \in B\} \in \mathcal{F}$ siempre que

$B \in \mathcal{B}(\mathbb{R})$. Decimos que la medida μ en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ es la distribución de X si

$$\mathbb{P}(X \in A) = \mu(A) = \int_A \mu(dx).$$

Observación. Equivalentemente se puede definir la función de distribución de X como la función $F : \mathbb{R} \longrightarrow [0, 1]$ tal que

$$F(x) = \mathbb{P}(X \leq x).$$

Definición A.0.3. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. $X : \Omega \longrightarrow \mathbb{R}$ es una variable aleatoria con función de distribución $\mathbb{P}X^{-1}$ en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Sea $g : \mathbb{R} \longrightarrow \mathbb{R}$ medible. Entonces,

$$E(g(X)) = \int_\Omega g(X(w))\mathbb{P}(dw)$$

es la esperanza de $g(X)$.

Definición A.0.4. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Un vector aleatorio es una función $X : \Omega \longrightarrow \mathbb{R}^n$ tal que para cualquier conjunto B en $\mathcal{B}(\mathbb{R}^n)$, se cumple que la imagen inversa $X^{-1}B$ es un elemento de $\mathcal{F}$.

Notemos que $X : \Omega \longrightarrow \mathbb{R}^n$ se puede escribir como $X = (X_1, ..., X_n)$. Por un resultado de teoria de la medida X es vector aleatorio si, y solo si, $X_i : \Omega \longrightarrow \mathbb{R}$ es variable aleatoria, para cada i.

Definición A.0.5. Sea $X = (X_1, ..., X_n)$ vector aleatorio en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$. La función

$$F(x_1, .., x_n) = \mathbb{P}(X_1 \leq x_1, ..., X_n \leq x_n)$$

es llamada la función de distribución conjunta de X.

Hay varios modos de convergencia en la teoría de probabilidades. Vamos a considerar algunos de ellos en nuestro análisis. Sea $\{X_n\}_{n=1}^{\infty}$ una sucesión de variables aleatorias en un espacio de probabilidad $(\Omega, \mathcal{F}, \mathbb{P})$ y sea X otra variable aleatoria en este mismo espacio, con distibución F.

Definición A.0.6. Sea F_n la función de distribución de X_n, $n \geq 1$. Decimos que X_n converge en distribución a X, en notación $X_n \xrightarrow{d} X$, si

$$F_n(x) \to F(x)$$

para x punto de continuidad de F.

Definición A.0.7. Decimos que $X_n \to X$ casi seguramente (c.s), si existe $N \in \mathcal{F}$ tal que $\mathbb{P}(N) = 0$ y $X_n(w) \to X(w)$ para $w \in N^c$.

Se suele usar la notacion $\mathbb{P}(X_n \to X) = 1$ y decir que $X_n \to X$ con probabilidad 1.

A.0.1. Borel-Cantelli

El Lema Borel-Cantelli es muy simple, pero sigue siendo la herramienta básica para demostrar una convergencia casi segura. Sea $(\Omega, \mathcal{F}, \mathbb{P})$ un espacio de probabilidad. Consideremos $\{A_n\}_{n=1}^{\infty}$ sucesión de eventos. Usaremos la notación $\{A_n, i.v.\}$ para $límsup A_n$, donde $i.v$ significa infinitas veces.

Teorema A.0.1. (Borel-Cantelli) *Si se cumple que* $\sum_n \mathbb{P}(A_n) < \infty$, *entonces* $\mathbb{P}(A_n, i.v) = 0$.

Teorema A.0.2. (Borel-Cantelli) *Si* $\{A_n\}_{n=1}^{\infty}$ *es una sucesión de eventos independientes y* $\sum_n \mathbb{P}(A_n) = \infty$, *entonces* $\mathbb{P}(A_n, i.v) = 1$.

Sea $\{X_n\}_{n=1}^{\infty}$ sucesión de variables aleatorias en $(\Omega, \mathcal{F}, \mathbb{P})$ y sea X otra variable aleatoria en el mismo espacio. Recordemos que cuando se muestra $\{X_n \to X\} \in \mathcal{F}$ se obtiene que

$$\{X_n \to X\} = \bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left(\mid X_k - X \mid \leq \frac{1}{m} \right).$$

Proposición A.0.1. *Supongase que para toda* $\epsilon > 0$, $\sum_n \mathbb{P}(\mid X_n - X \mid \geq \epsilon) < \infty$, *entonces* $X_n \to X$ *con probabilidad* 1.

Demostración. Es suficiente demostrar que

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left(\mid X_k - X \mid \leq \frac{1}{m} \right) \right) = 1.$$

Sea $\epsilon > 0$. Por el teorema A,0,1 se tiene que el conjunto $\{\mid X_n - X \mid \geq \epsilon, i.v.\}$ tiene probabilidad 0. Como $líminf\{\mid X_n - X \mid < \epsilon\} = \{\mid X_n - X \mid \geq \epsilon, i.v.\}^c$ resulta que $\mathbb{P}\left(líminf\{\mid X_n - X \mid < \epsilon\}\right) = 1$. Por ser ϵ arbitrario, en particular se tiene que

$$\mathbb{P}\left(líminf\{\mid X_n - X \mid < \frac{1}{m}\} \right) = 1$$

para todo $m \in \mathbb{N}$. De aquí concluimos que,

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} \left(\mid X_k - X \mid \leq \frac{1}{m} \right) \right) = \mathbb{P}\left(\bigcap_{m=1}^{\infty} líminf\{\mid X_n - X \mid < \frac{1}{m}\} \right) = 1.$$

□

A.0.2. Ley de Grandes Números

En la teoría de la probabilidad, bajo el término genérico de ley de los grandes números se engloban varios teoremas que describen el comportamiento del promedio de una sucesión de variables aleatorias conforme aumenta su número de ensayos.

Teorema A.0.3. *Sea $(X_n)_{n\geq 1}$ una sucesión de variables aleatorias independientes a pares con igual distribución y supongamos que $\mathbb{E}|X_1| < \infty$. Entonces,*

$$\frac{X_1 + ... + X_n}{n} \to \mathbb{E}(X_1)$$

con probabilidad 1.

A.0.3. Teorema Central del Limite (TCL)

Es uno de los resultados fundamentales de la teoría de probabilidad clasica. Presentamos inicialmente el teorema central de límite para el caso de variables i.i.d. y luego consideramos el caso más general de sucesión independientes pero no idénticamente distribuidas, que se conoce como el Teorema de Lindeberg-Feller.

Teorema A.0.4. (TCL para v.a.i.i.d.)
Sea $(X_n, n \geq 1)$ una sucesión de v.a.i.i.d con $\mathbb{E}(X_n) = \mu$ y $Var(X_n) = \sigma^2$. Sea $N \sim \mathcal{N}(0,1)$ y $S_n = X_1 + ... + X_n$. Entonces,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N.$$

Ahora generalizamos el TCL al caso de sumandos que no tienen la misma distribución. Sea $(X_n, n \geq 1)$ una sucesión de v.a.i. no necesariamente idénticamente distribuidas, supongamos que X_k tiene distribución F_k y $E(X_k) = 0$, $Var(X_k) = \sigma_k^2$. Definimos $s_n = \sigma_1^2 + ... + \sigma_k^2$.

Definición A.0.8. La sucesión $(X_n, n \geq 1)$ satisface la condición de Linderberg si para todo t, cuando $n \to \infty$, se tiene que

$$\frac{1}{s_n^2} \sum_{k=1}^{n} \mathbb{E}(X_k^2 1_{|X_k/s_n|>t}) \to 0.$$

Teorema A.0.5. (TCL para v.a.i)
La condición de Linderberg implica

$$\frac{X_1 + ... + X_n}{s_n} \xrightarrow{d} \mathcal{N}(0,1).$$

Observemos que el teorema del límite central establece que, en algunas situaciones, cuando se agregan variables aleatorias independientes, su suma correctamente normalizada tiende hacia una distribución normal incluso si las variables originales en sí no son normalmente distribuidas. Esto es clave en la teoría de la probabilidad porque implica que los métodos probabilísticos y estadísticos que funcionan para las distribuciones normales pueden ser aplicables a muchos problemas que involucran otros tipos de distribuciones.

Apéndice B

Otras Identidades Utilizadas

En este apéndice se encuentran algunas de las desigualdades y fórmulas para matrices que se utilizan en este trabajo, las cuales pueden ser encontradas en [26], [27] y [28].

Proposición B.0.1. (Fórmula de Woodbury) *Sean A, B, C y D matrices con A, C cuadradas y B, D posiblemente rectangulares. Supongamos que A y C son invertibles. Entonces,*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}.$$

La identidad de Woodbury, llamada así por Max A. Woodbury, es comunmente utilizada en la teoría de matrices aleatorias gracias a su particularidad de encontrar la inversa de una matriz en terminos de otra matriz. Como se ve en esta tesis, esta fórmula es vital para la conclusión de la Sección 3.1.

Proposición B.0.2. (Desigualdad de Mill) *Sea X una variable aleatoria con distribución $\mathcal{N}(\mu, \sigma^2)$. Entonces, para cualquier $\epsilon > 0$, se mantiene que*

$$\mathbb{P}(|X - \mu| > \epsilon) \leq \frac{1}{2\pi} \frac{e^{\frac{-\epsilon^2}{2\sigma^2}}}{t}.$$

La desigualdad de Mill ha sido ampliamente utilizada en la teoría de probabilidad clásica para la conclusión del Teorema de Borel-Cantelli. Como se vio en esta tesis, puntualmente en la Proposición 3.1.2, la desigualdad de Mill fue clave para la conclusión mediante el teorema de Borel-Cantelli de los puntos **2,3** y **6**.

Proposición B.0.3. (Desigualdad función Exponencial) *Sea $x \geq 0$. Entonces,*

$$e^x \geq 1 + x + \frac{x^2}{2}.$$

La Proposición B.0.2 fue una herramienta fundamental para la aplicación del teorema de Borel-Cantelli en la Proposición 3.1.2.

Proposición B.0.4. (Desigualdad de Jensen para la Traza) *Sea f una función continua definida sobre un intervalo I y sean $m, n \in \mathbb{N}$. Si f es convexa, tenemos la siguiente desigualdad*

$$Tr\left(f\left(\sum_{i=1}^{n} A_i^* X_i A_i\right)\right) \leq Tr\left(\sum_{i=1}^{n} A_i^* f\left(X_i\right) A_i\right)$$

para todas $(X_1, ..., X_n)$ matrices de tamaño $m \times m$ autoadjuntas con espectro en I y $(A_1, ..., A_n)$ matrices de tamaño $m \times m$ tales que $\sum_{i=1}^{n} A_i^* A_i = Id$.

La desigualdad de Jensen para la traza ha sido a lo largo de la historia una de las principales desigualdades que involucran al funcional traza. Como se vio en la Proposición 3.2.2 de la Sección 3.2, fue de gran ayuda el uso de esta desigualdad.

Apéndice C

Código Python

El código computacional de la red neuronal artificial utilizada en esta tesis puede ser encontrado en los siguientes dos enlaces:
https://colab.research.google.com/drive/1AZRq7vsNyFjm4bYcnKJDOD8ZyAenOXXD
https://colab.research.google.com/drive/1Lrj7oYdYAJ8ZS2U1WW4jnUXSfi981ta3

```
# Uso de tensorflow


import tensorflow as tf
import keras
from keras.models import Sequential
from keras.layers import Dense, Activation
from keras.datasets import mnist
from keras.utils import np_utils
import numpy as np
import numpy.random as rd
import matplotlib.pyplot as plt
%matplotlib inline


# Datos:


def make_data(n,p,c1,c2):
  n1 = int(n*c1)
  n2 = int(n*c2)
  mu = np.c_[2.0,np.zeros((1,p-1))];
  X1 = np.repeat(-mu,n1,axis=0);
  X2 = np.repeat( mu,n2,axis=0);
  X = np.r_[X1,X2]+rd.randn(n,p);
  Y = np.r_[-np.ones((n1,1)),np.ones((n2,1))];
  P =rd.permutation(np.arange(n));
  return X[P,:], Y[P];
```

```
n = 512;
p = 256;
c1 = 0.5;
c2 = 0.5;
X_train, Y_train = make_data(n,p,c1,c2);
X_test,  Y_test  = make_data(n,p,c1,c2);


#Red neuronal:


Eloss = np.zeros((300,));
Eval_loss = np.zeros((300,));
Eacc = np.zeros((300,));
Eval_acc = np.zeros((300,));

T = 50;
for t in range(T):
  print(t);
  X_train_ruido, Y_train_ruido = make_data(n,p,c1,c2);

  # Create model:
  model = Sequential()
  model.add(Dense(1, input_shape=(p,),
  use_bias=False,
  kernel_initializer=keras.initializers.RandomNormal(mean=0.0,
  stddev=np.sqrt(0.1/p))))

  #Compile model:
  sgd = keras.optimizers.SGD(lr=0.01)
  model.compile(optimizer=sgd,
              loss='mean_squared_error',
              metrics=['acc'] )
  history = model.fit(X_train_ruido, Y_train_ruido,
  epochs=300, verbose=0, batch_size=n,
  validation_data=(X_test,Y_test));
  Eloss += np.array(history.history['loss'])/T;
  Eval_loss += np.array(history.history['val_loss'])/T;
  Eacc += np.array(history.history['acc'])/T;
  Eval_acc += np.array(history.history['val_acc'])/T;


# Para graficar:


def plot_loss(uno,dos, path):
    plt.style.use("ggplot")
```

```
    f,ax = plt.subplots(1,1)
    ax.plot(uno)
    ax.plot(dos)
    ax.set_ylim(0.0,1.0)
    plt.title("Model's training loss")
    ax.set_xlabel("Epoch #")
    ax.set_ylabel("Loss")
    ax.legend(['Train', 'Test'], loc='upper left')
    #ax.legend(['Train'], loc='upper left')
    plt.savefig(path)

def plot_acc(uno,dos, path):
    plt.style.use("ggplot")
    f,ax = plt.subplots(1,1);
    ax.plot(1.0-uno);
    ax.plot(1.0-dos);
    ax.set_ylim(0.0,1.0)
    #plt.plot(history.history['val_acc'])
    plt.title("Model's training acc")
    ax.set_xlabel("Tiempo")
    ax.set_ylabel("Tasa de clasififcación errónea")
    ax.legend(['Entrenamiento', 'Generalización'], loc='upper right')
    #ax.legend(['Train'], loc='upper right')
    plt.savefig(path)

# Plot loss and accuracy:
plot_acc(Eacc,Eval_acc, 'model_acc_p.png')
plot_loss(Eloss,Eval_loss, 'model_loss_p.png')

#Uso tensorflow:


import tensorflow as tf
import keras
from keras.models import Sequential
from keras.layers import Dense, Activation
from keras.datasets import mnist
from keras.utils import np_utils
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline


#Se carga MNIST:


(x_train, y_train), (x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0
```

```
#Utilizamos numeros 1 y 7
X_train, X_test = [], []
Y_train, Y_test = [], []
sig = 0.0;
for lab,mat in zip(y_train,x_train):
  if lab == 1 or lab == 7:
    Y_train.append(-1 if lab == 1 else 1)
    X_train.append(mat.flatten())
for lab,mat in zip(y_test,x_test):
  if lab == 1 or lab == 7:
    Y_test.append(-1 if lab == 1 else 1)
    X_test.append(mat.flatten())
X_train = np.array(X_train)
X_test = np.array(X_test)
Y_train = np.array(Y_train)
Y_test = np.array(Y_test)
X_1 = X_train[Y_train < 0,:]
X_7 = X_train[Y_train > 0,:]
mu_1 = np.mean(X_1,axis=0);
mu_7 = np.mean(X_7,axis=0);
mu   = 0.5*(mu_7-mu_1)
X_1 -= np.repeat(mu_1.reshape((1,-1)),X_1.shape[0],axis=0);
X_7 -= np.repeat(mu_7.reshape((1,-1)),X_7.shape[0],axis=0);
C_1 = np.matmul(X_1.T,X_1)/X_1.shape[0];
C_7 = np.matmul(X_7.T,X_7)/X_7.shape[0];
U_1,D_1,_ = np.linalg.svd(C_1);
U_7,D_7,_ = np.linalg.svd(C_7);
isqrt_D_1 = np.diag(1.0/(np.sqrt(D_1)));
isqrt_D_7 = np.diag(1.0/(np.sqrt(D_7)));
isqrt_C_1 = np.matmul(U_1,np.matmul(isqrt_D_1,U_1.T));
isqrt_C_7 = np.matmul(U_7,np.matmul(isqrt_D_7,U_7.T));
X_1 = np.matmul(X_1,isqrt_C_1)-np.repeat(mu.reshape((1,-1)),X_1.shape[0],axis=0)
X_7 = np.matmul(X_7,isqrt_C_7)+np.repeat(mu.reshape((1,-1)),X_7.shape[0],axis=0)
X_train[Y_train < 0,:] = X_1;
X_train[Y_train > 0,:] = X_7;
sig =1e-5*np.sqrt(np.mean(X_train**2)-np.mean(X_train)**2);


#Red Neuronal:


Eloss = np.zeros((300,));
Eval_loss = np.zeros((300,));
Eacc = np.zeros((300,));
Eval_acc = np.zeros((300,));

T = 100;
```

```
for t in range(T):
  print(t);
  X_train_ruido = np.zeros((784,784));
  Y_train_ruido = np.array(392*[-1.0,1.0]);
  for i in range(392):
    idx = np.random.randint(X_1.shape[0]);
    jdx = np.random.randint(X_7.shape[0]);
    X_train_ruido[2*i  ,:] = X_1[idx,:];
    X_train_ruido[2*i+1,:] = X_7[jdx,:];
  #X_train_ruido = np.copy(X_train);
  #Y_train_ruido = np.copy(Y_train);
  X_train_ruido += sig*np.random.randn(X_train_ruido.shape[0],
  X_train_ruido.shape[1]);

  # Create model:
  model = Sequential()
  model.add(Dense(1, input_shape=(784,),use_bias=False,
  kernel_initializer=keras.initializers.RandomNormal(mean=0.0,
  stddev=np.sqrt(0.1)/28.0)))

  # Compile model:
  sgd = keras.optimizers.SGD(lr=0.01)
  model.compile(optimizer=sgd,
              loss='mean_squared_error',
              metrics=['acc'] )
  history = model.fit(X_train_ruido, Y_train_ruido, epochs=300,
  verbose=0, batch_size=784, validation_data=(X_train,Y_train));
  Eloss += np.array(history.history['loss'])/T;
  Eval_loss += np.array(history.history['val_loss'])/T;
  Eacc += np.array(history.history['acc'])/T;
  Eval_acc += np.array(history.history['val_acc'])/T;

# Para predecir, creamos algún dato random:
idx = np.random.randint(X_test.shape[0]);
R_test = X_test[idx,:]
a = model.predict(R_test.reshape(1,784));
print(Y_test[idx],a)


#Para Graficar:


def plot_loss(uno,dos, path):
    plt.style.use("ggplot")
    f,ax = plt.subplots(1,1)
    ax.plot(uno)
    ax.plot(dos)
    ax.set_ylim(0.0,1.0)
```

```
    plt.title("Model's training loss")
    ax.set_xlabel("Epoch #")
    ax.set_ylabel("Loss")
    ax.legend(['Train', 'Test'], loc='upper left')
    #ax.legend(['Train'], loc='upper right')
    plt.savefig(path)


def plot_acc(uno,dos, path):
    plt.style.use("ggplot")
    f,ax = plt.subplots(1,1);
    ax.plot(1.0-uno);
    ax.plot(1.0-dos);
    ax.set_ylim(0.0,1.0)
    #plt.plot(history.history['val_acc'])
    plt.title("Model's training acc")
    ax.set_xlabel("Tiempo")
    ax.set_ylabel("Tasa de clasififcación errónea")
    ax.legend(['Entrenamiento', 'Generalización'], loc='upper right')
    #ax.legend(['Train'], loc='upper right')
    plt.savefig(path)

# Plot loss and accuracy:
plot_acc(Eacc,Eval_acc, 'model_acc.png')
plot_loss(Eloss,Eval_loss, 'model_loss.png')
```

Referencias

[1] J. Wishart: *Generalized product moment distribution in samples.* Biometrika 20A 32, 1928.

[2] E. P. Wigner: *On the statistical distribution of the widths and spacings of nuclear resonance levels.* Proc. Cam. Phil. Soc. 47 790, 1951.

[3] R. Couillet, M. Debbah: *Random Matrix Methods for Wireless Communications.* Cambridge University press, 2011.

[4] Z. Bai, J. W. Silverstein: *Spectral Analysis of Large Dimensional Random Matrices.* Springer, 2009.

[5] T. Tao: *Topics in Random Matrix Theory.* Notas no publicadas, 2010.

[6] E. P. Wigner: *Characteristic vectors of bordered matrices with infinite dimensions.* Ann. Math. 62 548, 1955.

[7] P. I. Davies, N. J. Higham: *Computing f(a)b for matrix functions f.* Technical Report 436, School of Mathematics, University of Manchester, December 2004.

[8] D. Voiculescu: *Operator Algebras and their Connections with Topology in Ergodic Theory.* Lecture Notes in Mathematics 1132, Springer Verlag pp 556-588, 1983.

[9] S. G. Bobkov, F. Götze, A. N. Tikhomirov: *On concentration of high-dimensional matrices with randomly signed entries.* Springer, 2009.

[10] J. A. Mingo, R. Speicher: *Free Probability and Random Matrices.* Springer, 2017.

[11] R. Speicher: *Multiplicative functions on the lattice of noncrossing partitions and free convolution.* Mathematische Annalen 298, 611-628, 1994.

[12] R. Speicher: *Free Probability and Random Matrices.* Preprint arXiv:1404.3393, 2014.

[13] J. A. Mingo: *Non-crossing annular pairings and the infinitesimal distribution of the goe.* Preprint arXiv:1808.02100, 2019.

[14] S. Belinschi, D. Shlyakhtenko: *Free Probability of Type B: Analytic Interpretation and Applications.* Amer. J. Math. 134, 193-234, 2012.

[15] P. Biane, F. Goodman, A. Nica: *Non-crossing Cumulants of Type B.* Trans. Amer. Math. Soc. 355, 2263-2303, 2003.

[16] C. Popa: *Limit theorems and S-transform in non-ommutative probability spaes of type B.* Preprint arXiv:0709.0011, 2007.

[17] A. M. Saxe, J. L. McClelland, S. Ganguli: *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks.* Preprint arXiv:1312.6120, 2013

[18] Z. Liao, R. Couillet: *The Dynamics of Learning: A Random Matrix Approach.* 35th International Conference on Machine Learning, 2018.

[19] Boyd, Vandenberghe: *Convex optimization.* Cambridge university press, 2004.

[20] M. S. Advani, A. M. Saxe: *High-dimensional dynamics of generalization error in neural networks.* Preprint arXiv:1710.03667, 2017.

[21] D. Jakubovitz, R. Giryes, D. Rodrigues: *Generalization Error in Deep Learning.* Preprint arXiv:1808.01174, 2019.

[22] R. Rojas: *Neural Networks A Systematic Introduction.* Springer, 1996.

[23] Goodfellow, Bengio, Courville: *Deep Learning.* MIT Press, 2016.

[24] Groux, Benjamin: *Asymptotic freeness for rectangular random matrices and large deviations for sample covariance matrices with sub-Gaussian tails.* Electron. J. Probab. 22, paper no. 53, 40 pp. doi:10.1214/17-EJP4326, 2017.

[25] S. I. Resnick: *A Probability Path.* Birkhauser, Boston • Basel • Berlin, 1999.

[26] Max A. Woodbury: *Inverting modified matrices.* Memorandum Rept. 42, Statistical Research Group, Princeton University, Princeton, NJ, 1950.

[27] V. V. Buldygin, V. Kozachenko: *Sub-Gaussian random variables.* Springer 1980.

[28] F. Hansen, G. K. Pedersen: *Jensen's Trace Inequality in Several Variables.* International Journal of mathematics 14, 667-681, arXiv:math/0303060, 2003.