



UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Matemática

Tesis de Licenciatura

Geometría de redes neuronales convolucionales lineales

Emilia Vayssier

Directora: Alicia Dickenstein

Fecha

Agradecimientos

El camino hasta escribir esta tesis fue largo y sinuoso. Muchas personas me acompañaron a lo largo de los años, espero poder transmitirles mi reconocimiento. A los que siguen a mi lado y a los que no, gracias por su apoyo.

Gracias a Alicia por aceptarme como alumna sin conocerme y dirigir esta tesis. Gracias por tu generosidad con tu tiempo y tu conocimiento, disfruté mucho el verte hacer matemática.

Gracias a Mercedes y a Santiago por aceptar ser jurados de esta tesis, por su lectura minuciosa y sus devoluciones tan atentas.

Gracias a mi familia por su aliento y cariño incondicional. Su presencia a lo largo de estos años me permitió recorrer este camino con confianza. Gracias Mamá por impulsarme cuando tenía miedo de rendir, por siempre estar para escucharme, por enseñarme a escribir. Gracias Papá por sacarme a andar en bici cuando me frustraba, por enseñarme a improvisar. Merci de toujours m'encourager. Les agradezco todo lo que le dedicaron a mi educación, toda la ayuda con mis tareas y por inculcarme algo tan valioso como el deseo de aprender. Gracias a mis hermanos por siempre sacarme risas cuando estamos juntos, por compartir mis enojos y alegrías. Gracias Ani por ser mi compañera de vida, por regalarme los días más lindos antes de mi defensa y el mejor festejo de recibida. Gracias por los abrazos y por toda la felicidad que me da tenerte cerca.

Agradezco tener tantos tíos, tías, primos y primas. Saber que siempre están me ayudó a elegir volver y seguir estudiando en la UBA. Gracias Vicky, Santi, Hele y Aman por acogerme durante la pandemia y darme un hogar, por todo el cariño y el aguante de estos años, por atravesar todos esos momentos entre risas. Gracias Marí, Chelo, Anita y Pe por todos los paseos para distraerme y las horas de escritura en Escobar. Gracias Juan, Cele y Ane por su hermosa presencia. Gracias MV, Ale y Paulita por siempre alentarnos a todos. Gracias a mis abuelos, Antonio y Sofía, los tuve presentes todos estos años y compartí, de alguna manera, este momento con ustedes.

Gracias a esos amigos que se vuelven familia. Agradezco a mis tías, las Ávidas Contertulias, que por momentos estaban más preocupadas por esta tesis que yo. A Andre, por recibirme en Sevilla en nuestro retiro de escritura, por todos los cafés y por las aventuras compartidas. A Mariana, Simón, Rita y Coco, por todos los anti-domingos, por el aliento constante, por las horas de escritura en su casa. A Re, Bebito y Paquita por nuestra familia ensamblada. A Lola, Tomi y Juan por acompañarme durante tantos años y crecer junto a mí. A Jime y Angelita por su escucha, sus sabios consejos y por

enseñarme tanto sobre la vida.

En todos estos años de carrera, lo que recuerdo con más cariño son las horas presenciales y virtuales de cursada, de estudio y de ayudantía junto a mis compañeras y compañeros. Agradezco a mis primeras amigas Jessi, Maru y Ale. A Juja y a sus padres, Coco y Maru, que me dieron además un hermoso tiempo de convivencia. A mi profe del CBC, Pablo Colombo, por sus valiosas palabras de aliento. A mi profe de la UPC, Pere Pascu, por su inmensa generosidad, su disposición y los cafés compartidos. A los amigos que conocí más tarde, Nico Agote, Chino, Facu, Teo, Rama, Lucas, Lu y Georgi.

A esos docentes increíbles que encima se vuelven amigos, Daro y Juani. A esos amigos de los que siempre se aprende: Guido, Agus, Lu, Maki, Leo, Nacho, Ceci y Dev. Gracias por enseñarme tanto como personas y como matemáticos, por los hermosos momentos compartidos y por siempre estar. Gracias por su dedicación, por su enorme labor docente y científica y por todo lo que le dan a la facultad. Gracias por recorrer este camino junto a mí, no hubiese sido lo mismo sin ustedes.

A Sol, por tu presencia radiante en mi vida, que hace que todo sea mejor. Gracias por todo lo compartido que es no numerable: las horas de videollamada, las risas, los llantos, los abrazos, los mates. Gracias por escucharme, por cuidarme y por cocinarme; por tu cariño y tu apoyo incondicionales que sacan lo mejor de mí. Le agradezco a la UBA haberte conocido.

A Ivo, el hermano que me dio la vida, porque no hay palabras suficientes para agradecerte. Cada una de estas páginas contiene horas junto a vos con música de fondo. Gracias por sentarte a pensar conmigo, por toda la compañía que nos hacemos y el aliento que me das. Gracias por sostenerme, por ayudarme a disfrutar del proceso y dejarme los mejores recuerdos de escritura. No sé cómo hubiera terminado este trabajo sin vos.

Agradezco a mis amigos y compañeros de Di Tella: Romi, Vicky, Santi, Nico, Lupe, Nahue, Chichi, Lara y Marian. Gracias por todo su apoyo cotidiano, por hacerme ir a trabajar con una sonrisa, por los cafés antes de arrancar. Su presencia y cuidados me ayudaron a sostener este proceso. Agradezco a mis alumnos a lo largo de estos años por interesarse por mí y porque también aprendo de ellos.

Gracias a la UBA por la formación, la forma de ver la vida y la red que me dio. La educación es un privilegio y compartir conocimiento es de las actividades más bellas y valiosas que podemos realizar así que quiero agradecer especialmente la enorme e incansable labor de los docentes y no docentes que se encargan de sostener el funcionamiento de la educación pública, para que todos y todas podamos acceder a ella. Fueron años muy preciados.

Índice general

Introducción	1
1. Redes convolucionales	5
1.1. Definiciones generales	5
1.2. Relación con el sistema de visión	13
2. Polinomios, discriminantes y raíces múltiples	21
2.1. Polinomios de grado tres	23
2.2. Polinomios de grado cuatro	28
3. El espacio de funciones	37
3.1. Arquitecturas <i>filling</i>	38
3.2. Matrices convolucionales de salto 1	40
3.2.1. Identificación polinomial	41
3.2.2. Caracterización del espacio de funciones	44
3.3. Generalización a saltos más grandes	53
3.3.1. Identificación polinomial	54
3.3.2. Caracterización del espacio de funciones	56
4. Optimización	61
4.1. Puntos críticos	63
4.2. Dinámicas de entrenamiento	80
4.3. Optimización de la función de error cuadrático	85
4.3.1. La función de error cuadrático	85
4.3.2. Puntos críticos	88
4.3.3. Un ejemplo	90
A. Teorema de Sturm	93

Introducción

En los últimos años, las redes neuronales ganaron gran popularidad por su capacidad de llevar a cabo tareas que los modelos computacionales previos no lograban. En el caso de las redes neuronales convolucionales, su punto fuerte es el reconocimiento de patrones que les permite identificar y clasificar objetos o señales en sistemas de voz y de visión [1]. El rango de usos es muy amplio, desde la predicción por parte de las compañías eléctricas de la demanda energética al diagnóstico por imágenes en Medicina.

Lo que distingue a las redes neuronales de otros modelos es su estructura en capas inspirada en la estructura interconectada de las neuronas en el cerebro.

De [5], se tiene la siguiente definición

Definición 0.0.0.1. Sean \mathcal{X} el conjunto de elementos de entrada, \mathcal{Y} el conjunto de elementos de salida y \mathcal{W} el espacio de parámetros.

Un *modelo paramétrico de machine learning* es una función

$$\begin{aligned} f : \mathcal{W} \times \mathcal{X} &\rightarrow \mathcal{Y} \\ (\theta, x) &\mapsto f_{\theta}(x) = y \end{aligned}$$

que asocia un par (θ, x) de valores de parámetros y de entrada a un valor de salida y .

Por ejemplo, si se quiere diseñar un algoritmo de reconocimiento de caracteres, el modelo definirá una función $f(\theta, x)$ y buscará los valores de los parámetros θ que permitan clasificar correctamente los caracteres presentados en las imágenes.

En el caso de las *redes neuronales convolucionales*, dados parámetros θ , la función f_{θ} se describirá como una composición de convoluciones. Una convolución es una transformación lineal con una estructura particular que definimos en (1.1.0.1).

Definición 0.0.0.2. Sean \mathcal{X} el conjunto de elementos de entrada, \mathcal{Y} el conjunto de elementos de salida y \mathcal{W} el espacio de parámetros. Una *red neuronal* de profundidad L es una función

$$\begin{aligned} f : \mathcal{W} \times \mathcal{X} &\rightarrow \mathcal{Y} \\ f(\theta, x) &= (f_L \circ \sigma \circ \dots \circ f_2 \circ \sigma \circ f_1)(x). \end{aligned}$$

Se denomina *capa* a cada función f_i que dependerá de los parámetros θ . La función σ se denomina *función de activación* y puede no ser lineal.

En el caso de una *red neuronal convolucional*, todas las capas son convoluciones y la

función de activación es la identidad. De esta forma, si llamamos W_1, \dots, W_L a las matrices de cada convolución, como en la Definición 1.1.0.1 del Capítulo 1,

$$f(\theta, x) = W_L \dots W_1 x.$$

La última capa se denomina *capa de salida* y es la que devuelve la imagen de la función.

Cada elección de parámetros θ da lugar a una función. El conjunto de todas las funciones que genera una red al variar los parámetros se llama *espacio de funciones*. En la bibliografía, se lo denomina también *neuromanifold* [?].

Definición 0.0.0.3. El *espacio de funciones* de un modelo paramétrico de *machine learning* como en la Definición 0.0.0.1 es

$$\mathcal{M} := \{f_\theta : \mathcal{X} \rightarrow \mathcal{Y} : \theta \in \mathcal{W}\}.$$

Es la imagen del *mapa de parametrización*

$$\begin{aligned} \mu : \mathcal{W} &\rightarrow \mathcal{M} \\ \theta &\mapsto f_\theta \end{aligned}$$

que le asigna a los parámetros la función correspondiente.

Describir el espacio de funciones es un problema desafiante [5]. Diseños de redes muy sencillos dan lugares a complejas familias de funciones. Más aún, el mapa de parametrización es típicamente no lineal y no inyectivo: una misma función puede admitir diversas parametrizaciones. En [6], se define una identificación entre convoluciones y polinomios que permitirá describir al espacio de funciones como un subconjunto de un espacio de polinomios con condiciones sobre sus coeficientes.

Este trabajo trata algoritmos de *aprendizaje supervisado*. Para esto, es necesario un conjunto de datos de entrenamiento. Se trata de un conjunto finito de pares de elementos de entrada y de salida. Cada par consta de un elemento de entrada y del elemento de salida que queremos que la red le asigne. Por ejemplo, para diseñar un algoritmo de clasificación de caracteres, el conjunto de datos de entrenamiento puede consistir en imágenes de caracteres y una etiqueta con el caracter representado en cada imagen. El objetivo es que la red halle los valores de los parámetros que permitan asignar a cada imagen del conjunto de entrenamiento el caracter indicado y que estos valores permitan además la correcta clasificación de nuevas imágenes, no comprendidas en el conjunto de entrenamiento.

A partir de los datos de entrenamiento, definiremos una *función objetivo* \mathcal{L} que a cada valor de parámetros $\theta \in \mathcal{W}$ le asigne para cada elemento de entrada la diferencia entre el valor de salida de la red y el valor correcto provisto en los datos. Esta asignación puede también realizarse en el espacio de funciones. Dado un conjunto de entrenamiento, una *función de pérdida* ℓ le asigna a cada función para cada elemento de entrada la

diferencia entre el valor de salida de la red y el valor correcto provisto en los datos. La relación entre estas funciones se describe en la ecuación (4.0.0.2). Se plantean entonces dos problemas de optimización: hallar los valores de parámetros que minimizan este error y hallar la función del espacio de funciones que minimice el error.

El trabajo presentado se basa en el artículo [6] y estudia un tipo de red neuronal convolucional en particular: todas sus capas son convoluciones de salto 1 y los elementos de entrada son vectores. Los autores de [6] amplían su investigación a redes que admiten capas con mayores saltos en [7].

En el Capítulo 1, daremos las definiciones básicas relativas a redes neuronales convolucionales. Explicaremos su relación con el sistema de visión y el surgimiento de un modelo precursor, el *Neocognitron* [10], a partir de los descubrimientos de Hubel y Wiesel [9].

En el Capítulo 2, daremos las definiciones y resultados sobre espacios de polinomios para poder caracterizar el espacio de funciones correspondiente a una red neuronal convolucional. Daremos condiciones sobre los coeficientes de polinomios de coeficientes reales de grados 2, 3 y 4 para determinar la cantidad de raíces reales y su multiplicidad. Esto nos permitirá dar ejemplos de espacios de funciones asociados a distintas redes.

En el Capítulo 3, veremos al espacio de funciones como un subconjunto de un espacio de matrices convolucionales. Demostraremos que se trata de conjunto semi-algebraico que describiremos con ecuaciones e inecuaciones sobre los coeficientes de una matriz convolucional genérica. Daremos condiciones para determinar cuándo el espacio de funciones coincide con el espacio de matrices que lo contiene.

En el Capítulo 4, analizamos la existencia de puntos críticos para funciones objetivo y el impacto de la parametrización en su búsqueda. Damos una relación entre los puntos críticos del espacio de funciones y los del espacio de parámetros. Aplicamos los teoremas demostrados para hallar puntos críticos de la función de error cuadrático y comprobamos la convergencia del algoritmo de descenso del gradiente a estos puntos.

1. Redes convolucionales

1.1. Definiciones generales

En este trabajo, estudiaremos un tipo de red lineal llamada convolucional. Son muy eficaces para identificar patrones en imágenes con el fin de reconocer objetos. Se utilizan típicamente para clasificar imágenes, datos de audio y señales [3]. Para entender lo que la diferencia de otras redes, debemos comprender la operación de convolución.

En el contexto de los algoritmos de *machine learning*, la convolución es una operación matemática entre matrices. Una de sus entradas es el *filtro*, cuyas componentes son los parámetros que la red buscará ajustar. Comencemos con un ejemplo de convolución entre dos matrices.

Consideramos el filtro

$$\begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$$

y la matriz

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

Si identificamos una entrada de valor ‘0’ con un píxel negro y una entrada de valor ‘1’ con un píxel blanco, esta matriz codifica la imagen de una línea vertical y otra horizontal en cruz.

La operación de convolución dará como resultado una matriz de 3×3 . Para calcular el valor de la primera entrada, posicionamos el filtro en la esquina superior izquierda de la matriz de manera que se encuentre completamente dentro como en la Figura 1.1 de la izquierda y luego realizamos la multiplicación término a término de los coeficientes del filtro y las entradas de la matriz que se encuentran superpuestos. El valor de la entrada es la suma de estos productos:

$$1 \times 0 + 0 \times 0 + (-1) \times 1 + 1 \times 0 + 0 \times 0 + (-1) \times 1 + 1 \times 1 + 0 \times 1 + (-1) \times 1 = -2.$$

Para calcular la entrada siguiente, se desplaza el filtro una posición a la derecha siempre que se pueda disponer el filtro completamente dentro de la imagen. En este ejemplo,

1. Redes convolucionales

$$\begin{bmatrix} 0^1 & 0^0 & 1^{-1} & 0 & 0 \\ 0^1 & 0^0 & 1^{-1} & 0 & 0 \\ 1^1 & 1^0 & 1^{-1} & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0^1 & 0^0 & 1^{-1} & 0 & 0 \\ 1^1 & 1^0 & 1^{-1} & 1 & 1 \\ 0^1 & 0^0 & 1^{-1} & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Figura 1.1.: El recuadro naranja señala la posición del filtro sobre la matriz. Al lado de cada entrada de la matriz, se encuentra en naranja el coeficiente del filtro por el cual se multiplicará.
La imagen de la izquierda corresponde al cálculo de la primera entrada mientras que la de la derecha corresponde al cálculo de la cuarta entrada.

al finalizar la tercera iteración, ya no queda margen para desplazarse a la derecha. Por lo que se posiciona el filtro en el borde izquierdo una posición por debajo como en la Figura 1.1 de la derecha.

La matriz final obtenida es

$$\begin{bmatrix} -2 & 0 & 2 \\ -2 & 0 & 2 \\ -2 & 0 & 2 \end{bmatrix}.$$

Consideremos ahora las entradas con valores positivos como píxeles blancos y los no positivos como píxeles negros. La matriz resultante codifica una imagen de fondo negro con una única línea vertical. Intuimos entonces que este tipo de filtro detecta la presencia de líneas verticales. En [4], se realiza la convolución de este filtro con una imagen más compleja que consta de rectas con distintas orientaciones. En la imagen final solo quedan las rectas verticales. Las imágenes están dispuestas en la Figura 1.2.

Nos centraremos en la operación de convolución entre vectores.

Ejemplo 1.1.0.1. Consideremos el filtro w y el vector x dados por

$$w = [-1, 0, 1], x = [1, 3, 0, 2, -1].$$

La convolución de estos dos vectores resultará en otro vector de tres coordenadas. El filtro w recorre al vector x como se ilustra en la Figura 1.3. El cálculo de las coordenadas del vector final es

$$\begin{aligned} & [(-1) \times 1 + 0 \times 3 + 1 \times 0, (-1) \times 3 + 0 \times 0 + 1 \times 2, (-1) \times 0 + 0 \times 2 + 1 \times (-1)] \\ & = [-1, -1, -1]. \end{aligned}$$

A lo largo del trabajo, llamaremos convolución a un tipo de transformación lineal que definimos a continuación.



Figura 1.2.: A la izquierda, la imagen original. A la derecha, la imagen resultante de la convolución con el filtro.

$$\begin{array}{ccc} \begin{array}{ccc} -1 & 0 & 1 \\ 1 & 3 & 0 \end{array} & \begin{array}{ccc} -1 & 0 & 1 \\ 1 & 3 & 0 \end{array} & \begin{array}{ccc} -1 & 0 & 1 \\ 1 & 3 & 0 \end{array} \\ \begin{array}{cc} 2 & -1 \end{array} & \begin{array}{cc} -1 \end{array} & \begin{array}{cc} -1 \end{array} \end{array}$$

Figura 1.3.: El recuadro naranja señala la posición del filtro sobre el vector. Por encima de las coordenadas del vector se encuentran los coeficientes del filtro con los que se realizará la multiplicación.

Definición 1.1.0.1. Sean $s, k, d_0, d_1 \in \mathbb{N}$ relacionados por

$$sd_1 - d_0 = s - k. \quad (1.1.0.1)$$

Una *convolución* es una transformación lineal $\alpha : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_1}$ que tiene asociado un vector $w = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$ fijo, su *filtro*, y el entero positivo s que llamamos *salto*. Las coordenadas c_i del filtro se denominan *pesos* y el entero k indica su tamaño. Si $x \in \mathbb{R}^{d_0}$, la convolución α se expresa como

$$(\alpha(x))_i = \sum_{j \in [k]} c_j \cdot x_{is+j}, \quad (1.1.0.2)$$

con $0 \leq i \leq d_0 - 1$ y $[k] = \{0, 1, \dots, k-1\}$.

La matriz de α en la base canónica es de la forma

$$W = \begin{bmatrix} c_0 & c_1 & \dots & \dots & c_{k-1} & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & c_0 & c_1 & \dots & c_{k-1} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 0 & c_0 & c_1 & \dots & c_{k-1} \end{bmatrix}. \quad (1.1.0.3)$$

La fila i -ésima comienza con $s(i-1)$ ceros.

La relación (1.1.0.1) entre los tamaños de la entrada d_0 , de la salida d_1 , del filtro k y el

1. Redes convolucionales

salto s garantizan la buena definición de la transformación lineal.

Definimos además un mapa φ que toma una matriz convolucional $W \in \mathbb{R}^{d_1 \times d_0}$ y le asigna su filtro $w \in \mathbb{R}^k$

$$\varphi_{(d_0, d_1), k} : W \mapsto w. \quad (1.1.0.4)$$

A lo largo del trabajo lo denotaremos φ sin especificar las dimensiones de la matriz de entrada ni el tamaño de su filtro.

Podemos decir que una convolución actúa sobre un vector x recorriéndolo: se posa sobre una coordenada x_i del vector original, toma la suma ponderada de las k coordenadas siguientes y luego avanza s coordenadas para repetir el procedimiento en la coordenada x_{i+s} . El salto marca la velocidad de recorrido. En el Ejemplo 1.1.0.1, el filtro tiene tamaño $k = 3$ y coordenadas $(1, 0, -1)$, el salto es $s = 1$ e indica que el filtro se desplaza un lugar en cada iteración al recorrer el vector.

En este mismo ejemplo hemos denotado los vectores con corchetes pero a lo largo del trabajo lo haremos con paréntesis.

Ejemplo 1.1.0.2. Describimos una convolución $\alpha : \mathbb{R}^7 \rightarrow \mathbb{R}^3$ con filtro $w = (-1, 2, 4) \in \mathbb{R}^3$ de tamaño $k = 3$ y salto $s = 2$. Dado un vector de entrada $x \in \mathbb{R}^7$, el vector de salida es

$$\alpha(x) = \begin{bmatrix} -x_0 + 2x_1 + 4x_2 \\ -x_2 + 2x_3 + 4x_4 \\ -x_4 + 2x_5 + 4x_6 \end{bmatrix}.$$

Su matriz en la base canónica es

$$\begin{bmatrix} -1 & 2 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 2 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 2 & 4 \end{bmatrix}.$$

Damos un ejemplo más general de una convolución.

Ejemplo 1.1.0.3. Si se toma un filtro w de tamaño $k = 4$, como salto $s = 1$ y una entrada en \mathbb{R}^6 , podemos calcular el tamaño de la salida a partir de (1.1.0.1) como

$$\frac{6 - 4}{1} + 1 = 3.$$

La matriz en base canónica de la convolución es

$$\begin{bmatrix} w_0 & w_1 & w_2 & w_3 & 0 & 0 \\ 0 & w_0 & w_1 & w_2 & w_3 & 0 \\ 0 & 0 & w_0 & w_1 & w_2 & w_3 \end{bmatrix}.$$

Observación 1.1.0.1. Las matrices como en la Definición 1.1.0.3 se conocen como matrices de Toeplitz generalizadas. En este trabajo, las llamaremos *matrices convolucionales*.

Una matriz de Toeplitz es una matriz cuyas diagonales descendientes (de izquierda a derecha) tienen elementos constantes. Esto sucede con convoluciones de salto $s = 1$ como en el Ejemplo 1.1.0.3.

En el ámbito computacional, se llama convolución a la operación descrita al comienzo de esta sección. Se realiza entre vectores multidimensionales. Es la misma operación que está descrita en el Ejemplo 1.1.0.1 entre dos vectores. En cambio, en el marco de este trabajo, cada vez que se utilice la denominación de convolución hará referencia a una transformación lineal como la descrita en la Definición 1.1.0.1. Esta transformación lineal realiza la operación de convolución entre su filtro, un vector fijo, y otro vector que recibe como entrada.

Definimos a continuación una red lineal convolucional, el objeto de estudio de este trabajo.

Definición 1.1.0.2. Sea $L \in \mathbb{N}$, consideramos $\alpha_1, \dots, \alpha_L$ convoluciones definidas como en 1.1.0.1.

Cada convolución $\alpha_l : \mathbb{R}^{d_{l-1}} \rightarrow \mathbb{R}^{d_l}$ tiene asociado un tamaño de entrada $d_{l-1} \in \mathbb{N}$, un tamaño de salida $d_l \in \mathbb{N}$, un filtro $w^l = (w_0^l, \dots, w_{k_l-1}^l)$ con su tamaño $k_l \in \mathbb{N}$ y un salto $s_l \in \mathbb{N}$. Estas cantidades se relacionan como en (1.1.0.1)

$$d_l s_l - d_{l-1} = s_l - k_l. \quad (1.1.0.5)$$

Guardamos esta información en vectores que denotaremos $\mathbf{d} = (d_0, \dots, d_L)$, $\mathbf{k} = (k_1, \dots, k_L)$ y $\mathbf{s} = (s_1, \dots, s_L)$. La tupla $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ constituye la *arquitectura de la red*. Una *red lineal convolucional* con arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ es una familia de transformaciones lineales

$$\{f : \mathbb{R}^{d_0} \rightarrow \mathbb{R}^{d_L} / f(x) = (\alpha_L \circ \dots \circ \alpha_1)(x)\}. \quad (1.1.0.6)$$

Cada convolución α_i es una *capa* de la función f . Nos referimos a la cantidad L de capas como a la *profundidad* de la red. Llamaremos *capa de entrada* a α_1 y *capa de salida* a α_L .

Llamamos *espacio de parámetros* al conjunto $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ de vectores de filtros.

La matriz en base canónica $\overline{W} \in \mathbb{R}^{d_L \times d_0}$ de una transformación lineal f de la red se obtiene como el producto de las matrices convolucionales $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ de la Definición 1.1.0.3 asociadas a cada capa

$$\overline{W} = W_L W_{L-1} \dots W_2 W_1.$$

Una red lineal convolucional es un conjunto de transformaciones lineales con una estructura particular determinada por su arquitectura. Cada función de la red es una composición de convoluciones con tamaños de entrada, de salida, de filtro y saltos fijados por la arquitectura. Se obtienen las distintas transformaciones lineales que conforman la red al variar el valor de los pesos del filtro de cada capa. Estos pesos son los parámetros que ajustará el modelo computacional para obtener un mejor desempeño en la tarea para la que fue diseñado. Esto se ampliará en el Capítulo 4.

1. Redes convolucionales

Observación 1.1.0.2. Si se conocen el tamaño de la entrada de la red d_0 , el tamaño de la salida de la red d_L y los vectores de saltos $\mathbf{s} = (s_1, \dots, s_L)$ y de tamaños de filtro $\mathbf{k} = (k_1, \dots, k_L)$, la relación 1.1.0.5

$$d_l = \frac{d_{l-1} - k_l}{s_l} + 1$$

permite deducir los tamaños de entrada y de salida de las capas intermedias.

Damos un ejemplo a continuación.

Ejemplo 1.1.0.4. Si la arquitectura de la red es $\mathbf{d} = (5, 3, 2)$, $\mathbf{k} = (3, 2)$ y $\mathbf{s} = (1, 1)$, cada transformación lineal es una composición de dos convoluciones $\alpha_1 : \mathbb{R}^5 \rightarrow \mathbb{R}^3$ y $\alpha_2 : \mathbb{R}^3 \rightarrow \mathbb{R}^2$. El filtro de la primera es $w^1 = (w_0^1, w_1^1, w_2^1)$ de tamaño $k_1 = 3$ y el de la segunda es $w^2 = (w_0^2, w_1^2)$ de tamaño $k_2 = 2$. Ambas tienen salto $s_1 = s_2 = 1$. Un elemento de la red $f : \mathbb{R}^5 \rightarrow \mathbb{R}^2$ viene dado por $f(x) = (\alpha_2 \circ \alpha_1)(x)$ y calculamos su matriz en base canónica $\overline{W} \in \mathbb{R}^{5 \times 2}$ como el producto

$$\overline{W} = \begin{bmatrix} w_0^2 & w_1^2 & 0 \\ 0 & w_0^2 & w_1^2 \end{bmatrix} \begin{bmatrix} w_0^1 & w_1^1 & w_2^1 & 0 & 0 \\ 0 & w_0^1 & w_1^1 & w_2^1 & 0 \\ 0 & 0 & w_0^1 & w_1^1 & w_2^1 \end{bmatrix}.$$

Elijamos algunos valores para los pesos para ver qué forma tiene una transformación lineal particular de la red. Si $w^1 = (1, 0, -1)$ y $w^2 = (2, 1)$,

$$\overline{W} = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -2 & -1 & 0 \\ 0 & 2 & 1 & -2 & -1 \end{bmatrix}.$$

Podemos notar en el ejemplo que la transformación lineal resultante es nuevamente una convolución con filtro $w = (2, 1, -2, -1)$ de tamaño $k = 4$ y salto $s = 1$.

En general, al componer dos convoluciones cualesquiera, el resultado también es una convolución. Podemos calcular su filtro, el tamaño de su filtro y su salto a partir de los filtros y saltos de las convoluciones compuestas.

Más precisamente, si se consideran dos convoluciones $\alpha_1 : \mathbb{R}^{d_0} \mapsto \mathbb{R}^{d_1}$ y $\alpha_2 : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$ con filtros respectivos $w^1 \in \mathbb{R}^{k_1}$ y $w^2 \in \mathbb{R}^{k_2}$ que satisfacen (1.1.0.2), la composición viene dada por

$$\begin{aligned} (\alpha_2 \circ \alpha_1)(x)_i &= \sum_{j \in [k_1]} w_j^2 \cdot a_1(x)_{is_2+j} = \sum_{j \in [k_2]} w_j^2 \cdot \sum_{l \in [k_1]} w_l^1 \cdot x_{(is_2+j)s_1+l} \\ &= \sum_{j \in [k_2]} \sum_{l \in [k_1]} w_j^2 w_l^1 \cdot x_{is_2s_1+(js_1+l)} \\ &= \sum_{m \in [(k_2-1)s_1+k_1]} u_m \cdot x_{is_2s_1+m}. \end{aligned} \tag{1.1.0.7}$$

Como $0 \leq l \leq k_1 - 1$ y $0 \leq j \leq k_2 - 1$, el contador $m = js_1 + l$ puede tomar valores

$$\begin{aligned} l \leq js_1 + l \leq (k_2 - 1)s_1 + l &\implies 0 \leq js_1 + l \leq (k_2 - 1)s_1 + l \\ &\implies 0 \leq js_1 + l \leq (k_2 - 1)s_1 + k_1 - 1 \end{aligned}$$

El filtro u de la convolución resultante tiene tamaño $(k_2 - 1)s_1 + k_1$ y calculamos sus pesos como

$$u_m = \sum_{\substack{j \in [k_2], l \in [k_1] \\ js_1 + l = m}} w_j^2 w_l^1. \quad (1.1.0.8)$$

Si para algún valor de $m \in [(k_2 - 1)s_1 + k_1]$, no existen $j \in [k_2], l \in [k_1]$ tales que $m = js_1 + l$, el peso correspondiente u_m es nulo.

La convolución resultante tiene salto $s = s_1 s_2$.

En el siguiente resultado, generalizamos esta observación a una composición de L convoluciones. El salto de la convolución resultante es el producto de los saltos de cada convolución que se compone.

Proposición 1.1.0.1. *La composición de L convoluciones con filtros de tamaño $\mathbf{k} = (k_1, \dots, k_L)$ y saltos $\mathbf{s} = (s_1, \dots, s_L)$ resulta en una convolución con filtro de tamaño $k = k_1 + \sum_{l=2}^L (k_l - 1) \prod_{m=1}^{l-1} s_m$ y salto $s = \prod_{l=1}^L s_l$. Además, si todos los saltos s_i coinciden con el valor \bar{s} , entonces $k = d_0 - (d_1 - 1)\bar{s}$ y $s = \bar{s}^L$.*

Demostración. El resultado se demuestra por inducción en la cantidad de filtros L . El caso base $L = 2$ consiste en el desarrollo previo.

Se denota $\tilde{\alpha}$ a la composición de los primeros $L - 1$ filtros. Así,

$$(\alpha_L \circ \alpha_{L-1} \circ \dots \circ \alpha_1)(x)_i = (\alpha_L \circ \tilde{\alpha})(x)_i = \sum_{j \in [k_L]} w_j^L \cdot (\tilde{\alpha}(x))_{is_L + j}. \quad (1.1.0.9)$$

Por hipótesis inductiva, $\tilde{\alpha}$ es una matriz convolucional con tamaño de filtro $\tilde{k} = k_1 + \sum_{r=2}^{L-1} (k_r - 1) \prod_{t=1}^{r-1} s_t$, salto $\tilde{s} = \prod_{t=1}^{L-1} s_t$ y filtro \tilde{w} de forma que

$$(\tilde{\alpha}(x))_i = \sum_{l \in [\tilde{k}]} \tilde{w}_l \cdot x_{i\tilde{s} + l}.$$

Se puede entonces reescribir (1.1.0.9) como

$$\sum_{j \in [k_L]} w_j^L \cdot \left(\sum_{l \in [\tilde{k}]} \tilde{w}_l \cdot x_{(is_L + j)\tilde{s} + l} \right) = \sum_{j \in [k_L]} \sum_{l \in [\tilde{k}]} w_j^L \tilde{w}_l x_{is_L \tilde{s} + j\tilde{s} + l}$$

Llamamos $m = j\tilde{s} + l$ y calculamos los valores que puede tomar. Como $\tilde{s} = \prod_{t=1}^{L-1} s_t$, $0 \leq j \leq k_L - 1$ y $0 \leq l \leq k_1 + \sum_{r=2}^{L-1} (k_r - 1) \prod_{t=1}^{r-1} s_t - 1$, se tiene que

1. Redes convolucionales

$$\begin{aligned}
0 \leq m &\leq (k_L - 1) \prod_{t=1}^{L-1} s_t + k_1 + \sum_{r=2}^{L-1} (k_r - 1) \prod_{t=1}^{r-1} s_t - 1 \\
\implies 0 \leq m &\leq k_1 + \sum_{r=2}^L (k_r - 1) \prod_{t=1}^{r-1} s_t - 1.
\end{aligned}$$

La convolución resultante tiene tamaño de filtro $k_1 + \sum_{r=2}^L (k_r - 1) \prod_{t=1}^{r-1} s_t$ y salto $s = s_L \tilde{s} = \prod_{t=1}^L s_t$. □

Observación 1.1.0.3. Dada una red de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ y profundidad L , deducimos de este resultado que los elementos que conforman el espacio de funciones son a su vez convoluciones con tamaño de entrada d_0 , tamaño de salida d_L , tamaño de filtro

$$k = k_1 + \sum_{l=2}^L (k_l - 1) \prod_{m=1}^{l-1} s_m$$

y salto $s = \prod_{l=1}^L s_l$.

En el caso en que $\mathbf{s} = (1, \dots, 1)$, diremos que es una *red de salto 1*. Los elementos del espacio de funciones tienen tamaño de salto 1 y tamaño de filtro

$$k = k_1 + \sum_{l=2}^L (k_l - 1) = \sum_{l=1}^L k_l - (L - 1). \quad (1.1.0.10)$$

1.2. Relación con el sistema de visión

Las redes convolucionales son quizás la historia de mayor éxito de inteligencia artificial inspirada por la biología [2]. Si bien su desarrollo fue guiado por muchos campos, las características principales de su diseño se basan en la neurociencia. En esta sección, repasaremos sus orígenes en los descubrimientos de Hubel y Wiesel [9] y cómo dieron lugar a un modelo precursor, el *Neocognitron* [10]. Compararemos una versión simplificada del modelo de visión humana con la arquitectura más usual de redes convolucionales para el reconocimiento de objetos.

En su trabajo de 1962 [9], los científicos Hubel y Wiesel estudiaron las neuronas presentes en el cortex visual primario de gatos. Proyectaron estímulos lumínicos en ubicaciones precisas de una pantalla y registraron la respuesta de neuronas individuales. Esto les permitió distinguir entre neuronas simples y complejas.

Hallaron que, para cada neurona simple, la pantalla podía dividirse en regiones excitatorias, cuya iluminación provocaba una respuesta, y regiones inhibitorias, cuya iluminación suprimía la respuesta. Estas regiones eran antagonistas: iluminarlas en simultáneo no provocaba casi respuesta. Además, el cambio en la tasa de respuesta variaba según el tamaño del estímulo.

A partir de la disposición de estas regiones en la pantalla, pudieron predecir para cada neurona el estímulo estacionario más eficiente. Si bien distinguieron de distintos tipos, compartían una propiedad: las áreas excitatorias e inhibitorias se disponían lado a lado y las separaban líneas paralelas como puede observarse en la Figura 1.4 La orientación de los límites entre regiones era característico de cada célula.

Los estímulos estacionarios hallados más comunes fueron rectángulos de luz, podía haber varios siempre que fueran paralelos, y la división de la pantalla en dos regiones, una iluminada y otra oscura. En la Figura 1.4 se dan ejemplos oblicuos pero podían ser verticales u horizontales.

Dedujeron entonces que cada neurona simple era sensible a un tipo de estímulo sencillo, con su orientación característica, ubicado en una zona específica de la pantalla. En cambio, las neuronas complejas respondían a estímulos formados por patrones más complejos. Mantenían una respuesta consistente sin importar su ubicación en la pantalla. Propusieron un modelo en el que las neuronas complejas estaban conectadas a varias células simples que respondían al mismo tipo de estímulo en ubicaciones precisas.

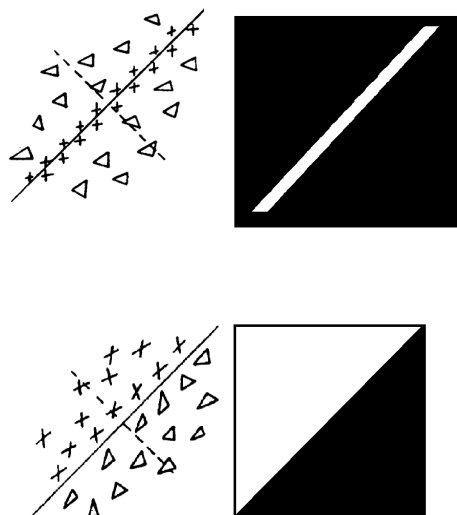


Figura 1.4.: A la izquierda, la disposición de las áreas excitatorias (representadas por un símbolo de cruz) y de las áreas inhibitorias (representadas por un símbolo de triángulo) separadas por un eje orientado. A la derecha, el estímulo estacionario más eficiente. La zona blanca representa la región iluminada.

Estos descubrimientos dirigieron la investigación a explicar el sistema de visión mediante modelos jerárquicos. Estos modelos se estructuran en capas dotadas de unidades de procesamiento, análogas a las neuronas. Cada unidad tiene una región del espacio visual a la que responde que llamamos *campo receptivo*. El campo receptivo de las unidades de un nivel de la jerarquía se construye combinando la salida de unidades de un nivel inferior. Después de algunas etapas de procesamiento, los campos receptivos pequeños y sensibles a estímulos simples se combinan para formar campos receptivos más grandes y sensibles a estímulos más complejos.

En un acercamiento simplificado al sistema de visión, al observar un objeto, la información fluye de la retina, a través de la región LGN, de las regiones del cortex visual V1 y V4 y finaliza su recorrido en el cortex prefrontal. A medida que se recorren estas regiones, los estímulos que excitan las neuronas se vuelven más complejos. En la región V1, las neuronas detectan rectas orientadas o bordes mientras que en el cortex prefrontal se han hallado neuronas sensibles a objetos complejos y hasta a caras [12]. Las neuronas se vuelven más tolerantes con respecto a la posición exacta del estímulo y a su escala. A su vez, su campo receptivo se vuelve mayor.

Los modelos jerárquicos son consistentes con el comportamiento estudiado de poblaciones de neuronas en las regiones nombradas, con los estudios de fMRI y la evidencia psicológica [12].

El comportamiento de las neuronas simples y complejas inspiró el diseño del *Neocognitron*, un modelo computacional que se aplicó para el reconocimiento de caracteres. Daremos una explicación simplificada de su funcionamiento para este caso.

El modelo recibe como entrada una imagen de un carácter y su objetivo es determinar cuál es. Se trata de una red compuesta por capas sucesivas. Cada capa contiene *células*. Son funciones que reciben como entrada los valores de salida de algunas células de la capa anterior, los procesan y emiten un valor de salida que recibirán algunas células de la capa siguiente. Dos células están *conectadas* si una recibe como entrada el valor de salida de la otra. Cada célula tiene un *campo receptivo* que consta de todas las células de la capa anterior cuyo valor de salida recibe como entrada.

En este contexto, utilizamos la denominación “capa” de forma distinta. Cada capa está formada de *planos*. Podemos imaginar que un plano es una grilla del tamaño de la imagen de entrada compuesto por células. Cada célula tiene como campo receptivo una región de un plano de la capa anterior. Dicha región es un entorno de la ubicación de la célula en su plano proyectada al plano anterior. Lo ilustramos en la Figura 1.6.

Existen dos tipos de células, las *simples* y las *complejas*, que emulan el comportamiento de las neuronas correspondientes. Cada célula simple detecta una característica particular en una ubicación específica del plano. Todas las células simples de un mismo plano detectan el mismo rasgo. Una célula compleja recibe la salida de varias células simples de un mismo plano con distintas ubicaciones. El resultado buscado es que una célula compleja responda al mismo estímulo sin importar su ubicación en la entrada, tal como lo hacen las neuronas complejas.

La red alterna una capa de células simples, denotada U_S , y una de células complejas U_C . Describimos esta arquitectura en la Figura 1.6.

Las células simples aprenderán qué característica detectar. A medida que se encuentren en capas más profundas, detectarán rasgos más complejos.

Veamos, de forma esquemática, cómo funciona esta arquitectura para detectar un

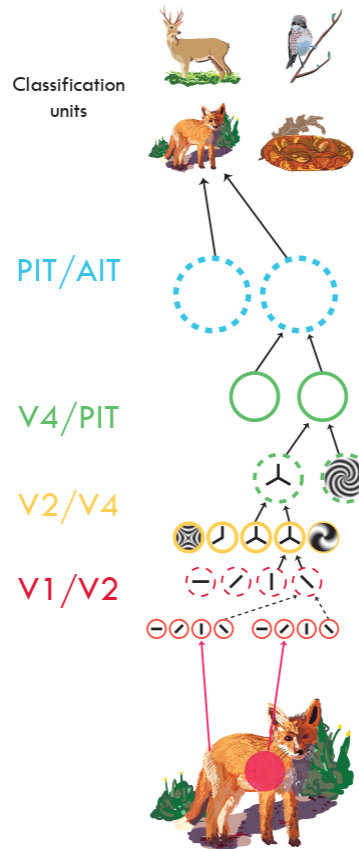


Figura 1.5.: Un esquema de un modelo jerárquico. Los acrónimos ‘PIT’ y ‘AIT’ designan zonas del cortex prefrontal y ‘V1’, ‘V2’ y ‘V4’ zonas del cortex visual.

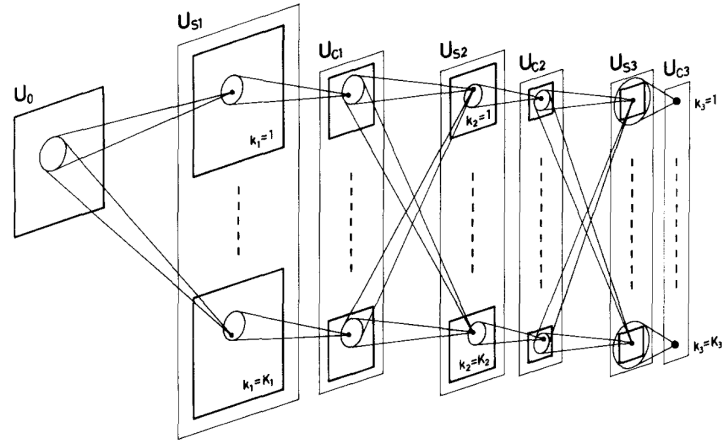


Figura 1.6.: Los rectángulos más grandes representan las capas y los más pequeños en su interior, los planos. Los círculos sobre cada plano indican los campos receptivos de las células de la capa siguiente. Podemos observar que el tamaño del campo receptivo aumenta con la profundidad de la red. La última capa, la de salida, tiene células cuyo tamaño de campo receptivo es toda la imagen de entrada.

carácter.

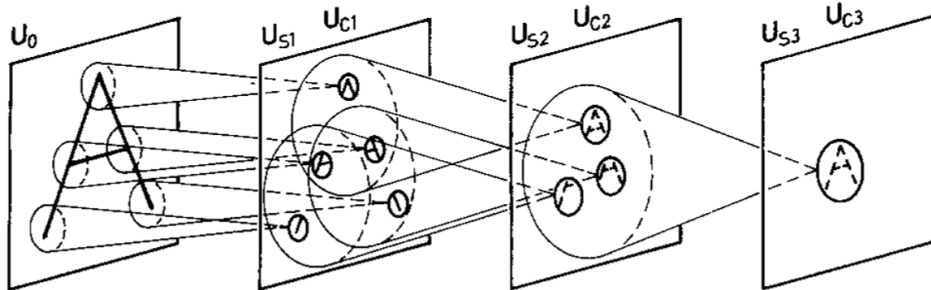


Figura 1.7.: El primer cuadrado U_0 es la imagen de entrada. En los cuadrados siguientes están dibujados los rasgos reconocidos por las capas indicadas. Los círculos indican los campos receptivos de las células.

En la Figura 1.7, U_0 es la imagen de entrada. En este ejemplo, es el carácter 'A'. El cuadrado siguiente ilustra los rasgos que detectan las células de los planos de la capas U_{S1} y U_{C1} . Son características simples: esquinas, líneas, intersecciones de líneas. A medida que se avanza en la red, el campo receptivo aumenta de tamaño. Las capas siguientes detectan rasgos más complejos que son combinaciones de los más simples detectados previamente.

Vemos en la Figura 1.8 los planos de las primeras capas.

El primer plano de U_{S1} consta de células simples que detectan esquinas orientadas hacia arriba. Cada célula simple detecta este patrón en la imagen en un entorno de su ubicación en el plano. Por ejemplo, notamos en la figura que solo se activan las células ubicadas en la región del plano que, superpuesto sobre la imagen, se encontraría la esquina superior del carácter 'A'. Al contrario, las células de las esquinas no se activan.

Podríamos decir que este patrón de esquina orientada hacia arriba es la salida de las células simples que recibirán las células complejas del plano correspondiente de la capa U_{C1} . Estas células detectan este mismo rasgo en distintas ubicaciones de su campo receptivo.

En la segunda capa, el campo receptivo de las células es más amplio. Reciben como entrada los rasgos distinguidos en la capa anterior y aprenden a detectar una combinación de ellos.

De la segunda capa U_{S2} , solo se ilustra un plano. El área cubierta por el campo receptivo de la célula presentada no abarca los extremos inferiores del carácter 'A'. Las células de este plano detectan una combinación de la esquina superior y los cruces de este carácter.

La capa final tendrá como campo receptivo toda la imagen y se habrá especializado en la detección de cada carácter.

Las redes convoluciones están directamente inspiradas en las nociones de neuronas simples y complejas y su arquitectura se asemeja al circuito 'LGN-V1-V2-V4-IT' descrito. Tienen sus raíces en el *Neocognitron* [11].

En las arquitecturas más comunes de redes convolucionales, podemos hallar capas convolucionales y capas de *pooling*.

En la Sección 1.1 describimos las capas convolucionales. Son análogas a las células simples. Podríamos pensar que los pesos de los filtros asociados codifican zonas excitatorias e inhibitorias. Un peso negativo codifica una zona inhibitoria mientras que uno positivo codifica una zona excitatoria. Esto, en los estratos iniciales, les permite detectar características simples como vimos en el Ejemplo 1.1 con el filtro que detecta las rectas verticales. En este trabajo, solo estudiaremos convoluciones cuya entrada es un vector. En general, en los modelos de procesamiento de imágenes, la entrada de la red es una

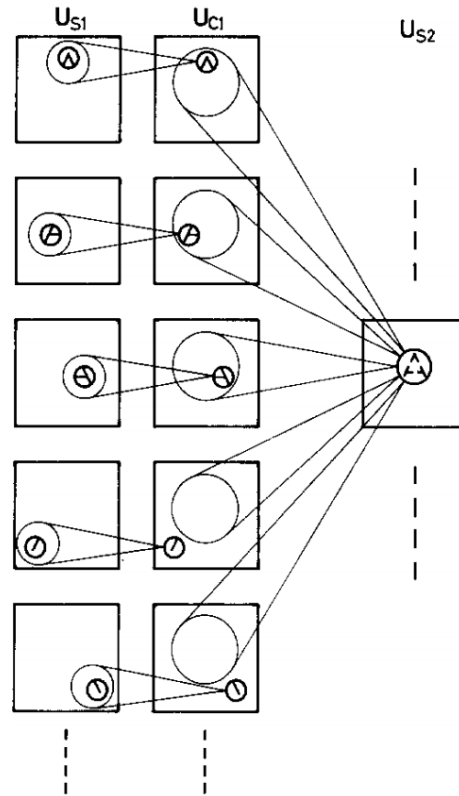


Figura 1.8.: Planos de las capas U_{S1} y U_{C1} con los rasgos detectados.

1. Redes convolucionales

matriz que codifica la imagen. Tras pasar por una capa de convolución, se obtiene un *mapa de activación*. Se trata de una copia de la imagen en la que solamente figura el rasgo detectado por la capa en las ubicaciones que se encontró este rasgo en la imagen original. Esto se asemeja a los planos del *Neocognitron* y tiene un eco en la estructura del cortex visual primario. Las neuronas simples tienen una distribución espacial retinotópica. Es decir, que las neuronas adyacentes tienen campos visuales contiguos. Su organización forma una representación topográfica del campo visual y tiene un comportamiento similar al del filtro de una convolución mientras recorre una imagen.

El objetivo de las capas de pooling es condensar la información que reciben de las capas convolucionales. Un ejemplo es la función **max pooling** que elige el elemento máximo de la región que abarca de su entrada. Las capas de pooling comparten características con las células complejas del *Neocognitron*, como la tolerancia a ciertas transformaciones del rasgo que detectan.

Las capas más profundas de una red son sensibles a estímulos más complejos y se asemejan a las neuronas halladas en el cortex prefrontal. ¿Cómo podemos comprender qué detecta cada capa de una red convolucional? Una estrategia es encontrar las imágenes del conjunto de entrenamiento que resulten en una mayor respuesta de la capa. Diremos que la activan. Otra estrategia es partir de una imagen de ruido aleatorio y, mediante procedimientos de optimización, obtener una imagen que genere la salida máxima de la capa. Podemos pensar que es buscar un mapa de activación óptimo para esa capa.

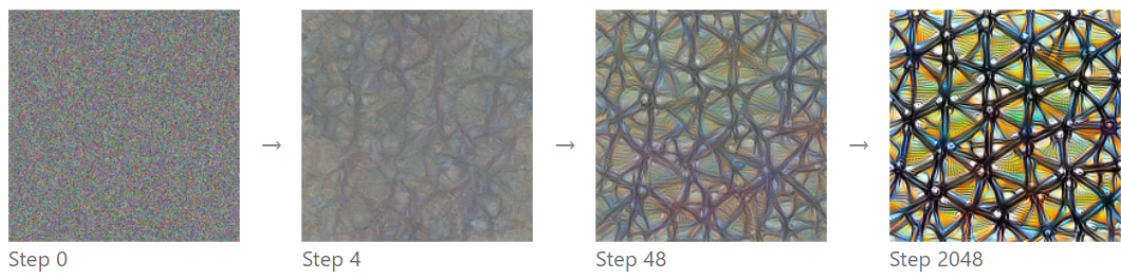


Figura 1.9.: Progresión de los pasos para hallar un mapa de activación óptimo para una capa. La imagen final es el patrón que activa esta capa.

En el artículo [13], se detallan estas técnicas y se proveen ejemplos de la red *Google-LeNet* entrenada con el conjunto de datos *ImageNet*. Mostraremos algunos de ellos a continuación.

En la Figura 1.10, comparamos las imágenes del conjunto de datos que generan una mayor activación con las que se obtuvieron mediante optimización. Al comparar las imágenes de la tercera columna, deducimos que la capa no está especializada en la detección de nubes sino en la de una textura “esponjosa”. La segunda columna ilustra una capa especializada en la detección de hocicos.

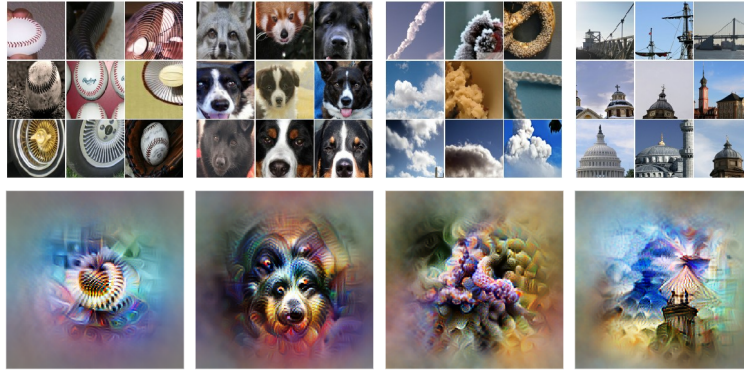


Figura 1.10.: Cada columna representa una capa. En la primera fila, las imágenes que generan la mayor activación. En la segunda, las imágenes que se obtuvieron mediante optimización.

Con estos ejemplos, ilustraremos la complejización de los rasgos detectados a medida que se avanza por la red.

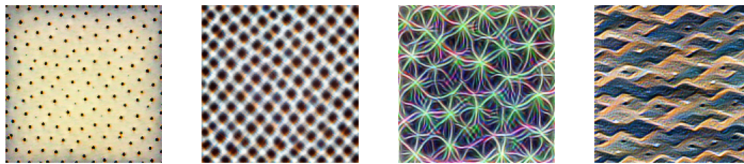


Figura 1.11.: En las capas iniciales, se detectan texturas o patrones básicos. Como el campo receptivo es pequeño, se muestra una composición en mosaico de los estímulos.

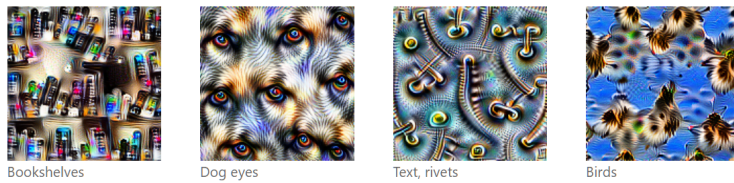


Figura 1.12.: En las capas siguientes, se distinguen patrones más complejos y hasta rasgos de objetos.

1. Redes convolucionales



Figura 1.13.: Las capas comienzan a discernir animales y objetos en un contexto. La distinción no es tan fina todavía. Se activan con conceptos o texturas similares, como en la segunda imagen que observamos pan y helado.

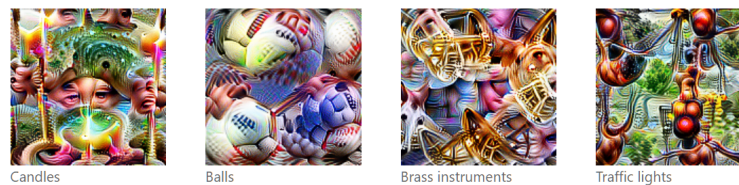


Figura 1.14.: Más adelante en la red, las capas son sensibles a conceptos semánticos diferenciados. Si bien la interpretación de las imágenes visualizadas se dificulta, aún se pueden distinguir los conceptos que las activan.



Figura 1.15.: En las capas más profundas, se pierden las nociones semánticas en la visualización.

2. Polinomios, discriminantes y raíces múltiples

Para poder comprender la estructura del espacio de funciones, trabajaremos con polinomios. Nos interesará encontrar criterios para determinar la cantidad de raíces reales de un polinomio dado y su multiplicidad. En este capítulo, listamos las definiciones y los resultados que necesitaremos en los capítulos siguientes.

Según el contexto, trabajaremos con polinomios en una variable de $\mathbb{R}[x]$ o con polinomios homogéneos de dos variables de $\mathbb{R}[x, y]$. Un *polinomio homogéneo* es un polinomio en que cada monomio tiene el mismo grado. Por ejemplo $p = x^4 + 2x^3y - x^2y^2 + 6xy^3 + y^4$ es un polinomio homogéneo de grado 4 en dos variables.

El *discriminante* de un polinomio $p = a_nx^n + a_{n-1}x^{n-1}y + \cdots + a_1xy^{n-1} + a_0y^n \in \mathbb{R}[x, y]$ es otro polinomio, con coeficientes enteros, en los coeficientes a_n, \dots, a_0 que se anula si y solamente si p tiene una raíz compleja doble distinta de $x = y = 0$. Damos la definición que figura en el Capítulo 4 de [14].

Definición 2.0.0.1. Si $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ es un polinomio mónico de grado n y x_1, \dots, x_n son sus raíces en \mathbb{C} repetidas acorde a su multiplicidad, podemos calcular el determinante como

$$\Delta = \prod_{n \geq i > j \geq 1} (x_i - x_j)^2.$$

Ejemplo 2.0.0.1. Si se tiene un polinomio $p(x) = x^2 + a_1x + a_0$, el discriminante se calcula como

$$\Delta = a_1^2 - 4a_0.$$

Observación 2.0.0.1. En el caso de polinomios en dos variables homogéneos y mónicos de grado n de la forma

$$p(x, y) = x^n + a_{n-1}x^{n-1}y + \cdots + a_0y^n$$

definiremos su discriminante como el discriminante detallado en 2.0.0.1 del polinomio homogéneo y mónico de grado n

$$p(x, 1) = p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0.$$

El signo del discriminante puede dar una idea de la cantidad de raíces reales del polinomio, como indica el siguiente resultado de [14]

2. Polinomios, discriminantes y raíces múltiples

Proposición 2.0.0.1. Sea $p \in \mathbb{R}[x]$ mónico, de grado n y con n raíces distintas en \mathbb{C} . Si denotamos t la cantidad de raíces reales de p , se tiene que

$$\begin{aligned}\Delta > 0 &\iff t \equiv n \pmod{4}, \\ \Delta < 0 &\iff t \equiv n - 2 \pmod{4}.\end{aligned}$$

Demostración. Sean y_1, \dots, y_t las raíces reales de p y $z_1, \bar{z}_1, \dots, z_s, \bar{z}_s$ sus raíces en $\mathbb{C} - \mathbb{R}$. Así, $n = 2s + t$.

Analizamos los distintos factores que pueden aparecer en el producto que define al discriminante:

- $(y_i - y_j)^2 > 0$, $1 \leq i < j \leq t$,
- $((y_i - z_j)(y_i - \bar{z}_j))^2 > 0$, $1 \leq i \leq t, 1 \leq j \leq s$
pues el conjugado de $y_i - z_j$ es $y_i - \bar{z}_j$,
- $((z_i - z_j)(z_i - \bar{z}_j)(\bar{z}_i - z_j)(\bar{z}_i - \bar{z}_j))^2 > 0$, $1 \leq i < j \leq s$
pues el conjugado de $z_i - z_j$ es $\bar{z}_i - \bar{z}_j$ y el conjugado de $z_i - \bar{z}_j$ es $\bar{z}_i - z_j$,
- $(z_i - \bar{z}_i)^2 < 0$, $1 \leq i \leq s$.

Luego, el signo del discriminante depende de los factores en que se resta una raíz compleja con su conjugada

$$\text{sign}(\Delta) = \text{sign}\left(\prod_{i=1}^s (z_i - \bar{z}_i)^2\right) = (-1)^s.$$

Como $t = n - 2s$, al mirar el resto en la división por 4, $t \equiv n - 2s \pmod{4}$,

$$\begin{aligned}t \equiv n \pmod{4} &\iff s \equiv 0 \pmod{4} \text{ o } s \equiv 2 \pmod{4} \\ &\iff s \equiv 0 \pmod{2} \\ &\iff \Delta > 0.\end{aligned}$$

y

$$\begin{aligned}t \equiv n - 2 \pmod{4} &\iff 2s \equiv 2 \pmod{4} \\ &\iff s \equiv 1 \pmod{4} \text{ o } s \equiv 3 \pmod{4} \\ &\iff s \equiv 1 \pmod{2} \\ &\iff \Delta < 0.\end{aligned}$$

□

Definición 2.0.0.2. Dado un entero positivo $n \in \mathbb{N}$, una *partición* es una forma de descomponerlo como una suma de enteros positivos. Para evitar repeticiones, una partición $(\lambda_1, \dots, \lambda_r)$ de n cumple que

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r \text{ y } n = \lambda_1 + \dots + \lambda_r.$$

Por ejemplo, las posibles particiones de $n = 3$ son (3) , $(2, 1)$ y $(1, 1, 1)$.

Definición 2.0.0.3. Dada una partición $\lambda = (\lambda_1, \dots, \lambda_r)$ de n , el *locus de raíces múltiples* $\Delta_\lambda \subset \mathbb{R}[x, y]_n$ es el conjunto de polinomios homogéneos de grado n que tienen r raíces complejas, no necesariamente distintas, con multiplicidades $\lambda_1, \dots, \lambda_r$. Es decir, Δ_λ es el conjunto de polinomios $P \in \mathbb{R}[x, y]_n$ que pueden escribirse como $P = Q_1^{\lambda_1} \dots Q_r^{\lambda_r}$ con cada $Q_i \in \mathbb{C}[x, y]_1$ una forma lineal compleja.

Ejemplo 2.0.0.2. $\Delta_{(2,1)}$ es el conjunto de polinomios de grado 3 que tienen una raíz doble y una simple. Estos polinomios pueden escribirse como

$$a(x - \alpha y)(x - \beta y)^2$$

con $a \in \mathbb{R}, \alpha, \beta \in \mathbb{C}$. Como en la definición no se exige que las r raíces sean distintas, la raíz simple puede coincidir con la raíz doble. Así, un polinomio de grado 3 con una raíz triple también está en $\Delta_{(2,1)}$.

En particular, los polinomios $(x - y)^3$ y $(x - y)^2(x - 3y)$ son elementos de $\Delta_{(2,1)}$ pero solo $(x - y)^3$ es un elemento de $\Delta_{(3)}$.

Nos interesa además distinguir entre raíces complejas y reales.

Definición 2.0.0.4. Un polinomio homogéneo $P \in \mathbb{R}[x, y]_n$ tiene un *patrón de multiplicidad de raíces reales*, abreviado *rrmp*, $(\rho|\gamma) = (\rho_1, \dots, \rho_r|\gamma_1, \dots, \gamma_c)$ si tiene r raíces reales distintas con multiplicidades ρ_1, \dots, ρ_r y c pares distintos de una raíz compleja y su respectiva conjugada con multiplicidades $\gamma_1, \dots, \gamma_c$. Es decir, que puede escribirse como

$$P = p_1^{\rho_1} \dots p_r^{\rho_r} q_1^{\gamma_1} \dots q_c^{\gamma_c},$$

con $p_i \in \mathbb{R}[x, y]_1$ y $q_j \in \mathbb{R}[x, y]_2$ irreducibles y linealmente independientes dos a dos. Cada factor q_j tiene dos raíces complejas conjugadas.

Ejemplo 2.0.0.3. Para polinomios de grado dos $p = ax^2 + bxy + cy^2$, el discriminante $\Delta = b^2 - 4ac$ distingue tres *rrmp* posibles:

11 0 dos raíces reales distintas:	$\Delta > 0$
2 0 una raíz real doble:	$\Delta = 0$
0 1 sin raíces reales:	$\Delta < 0$

En lo que sigue, ilustraremos los posibles *rrmp* de polinomios de grado tres y cuatro.

2.1. Polinomios de grado tres

Definición 2.1.0.1. Dado un polinomio de grado tres $p = ax^3 + bx^2y + cxy^2 + dy^3$ con coeficientes reales, se calcula el discriminante

$$\Delta = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd. \quad (2.1.0.1)$$

2. Polinomios, discriminantes y raíces múltiples

Se definen además las siguientes cantidades

$$\begin{aligned}\delta_1 &= 3ac - b^2, \\ \delta_2 &= 9ad - bc, \\ \delta_3 &= 3bd - c^2.\end{aligned}\tag{2.1.0.2}$$

Lema 2.1.0.1. *El discriminante de un polinomio de grado tres $p = ax^3 + bx^2y + cxy^2 + dy^3$ puede escribirse como*

$$\Delta = 4\delta_1\delta_3 - \delta_2^2.$$

Demostración. Consideramos un polinomio homogéneo $p(x, y) = Ax^3 + 3Bx^2y + 3Cxy^2 + Dy^3$. Respecto a los coeficientes del polinomio del enunciado, se tiene que $a = A$, $b = 3B$, $c = 3C$ y $d = D$.

La condición que define al discriminante es que se anula si y solamente si el polinomio tiene una raíz doble. Esta condición sobre los coeficientes también puede pensarse en términos del polinomio y de su derivada. Buscamos valores de A, B, C y D que permitan la existencia de una solución del siguiente sistema

$$\begin{aligned}p(x, y) &= Ax^3 + 3Bx^2y + 3Cxy^2 + Dy^3 = 0 \\ p_x(x, y) &= 3Ax^2 + 6Bxy + 3Cy^2 = 0 \\ p_y(x, y) &= 3Bx^2 + 6Cxy + 3Dy^2 = 0.\end{aligned}\tag{2.1.0.3}$$

Si bien pareciera que la elección de polinomios homogéneos agrega una ecuación, cabe observar que el polinomio es una combinación lineal de sus derivadas parciales

$$3p = xp_x + yp_y.$$

Esto permite elegir dos ecuaciones del sistema (2.1.0.3) con las que seguir trabajando. Nos quedamos con las de menor grado y, al sacar factor común 3, obtenemos un nuevo sistema

$$\begin{aligned}Ax^2 + 2Bxy + Cy^2 &= 0 \\ Bx^2 + 2Cxy + Dy^2 &= 0.\end{aligned}\tag{2.1.0.4}$$

Mediante la siguiente combinación lineal, podemos eliminar el término x^2

$$\begin{aligned}B(Ax^2 + 2Bxy + Cy^2) - A(Bx^2 + 2Cxy + Dy^2) &= 0 \\ \implies 2(B^2 - AC)xy + (BC - AD)y^2 &= 0.\end{aligned}$$

De manera simétrica, eliminamos el término y^2 . Al simplificar los factores x e y de cada ecuación resultante, obtenemos un sistema lineal

$$\begin{aligned}2(B^2 - AC)x + (BC - AD)y &= 0 \\ (AD - BC)x + 2(BD - C^2)y &= 0.\end{aligned}\tag{2.1.0.5}$$

Si el sistema (2.1.0.4) tiene solución, entonces también la tiene el sistema (2.1.0.5). Cada ecuación lineal puede resolverse y la condición de igualdad de las soluciones arroja

$$\frac{x}{y} = \frac{AD - BC}{2(B^2 - AC)} = \frac{-2(BD - C^2)}{AD - BC}.$$

Deducimos entonces que la condición que deben cumplir los coeficientes para que el sistema (2.1.0.3) admita una solución es

$$4(BD - C^2)(AC - B^2) - (AD - BC)^2 = 0.$$

Reescribimos la condición con los coeficientes del enunciado

$$4\left(\frac{b}{3}d - \frac{c^2}{9}\right)\left(a\frac{c}{3} - \frac{b^2}{9}\right) - (ad - \frac{bc}{9})^2 = 0.$$

Multiplicar la ecuación anterior por 81 nos permite escribir al determinante como

$$\Delta = 4(3bd - c^2)(3ac - b^2) - (9ad - bc)^2,$$

donde reconocemos las expresiones de δ_1, δ_2 y δ_3 . □

Se pueden dar condiciones en términos de Δ y de δ_1 para caracterizar los posibles *rrmp*.

Proposición 2.1.0.1. *Se distinguen cuatro posibles rrmp para un polinomio de grado tres $p = ax^3 + bx^2y + cxy^2 + dy^3$*

111 0 tres raíces reales distintas:	$\Delta > 0$
12 0 una raíz real simple y una raíz real doble:	$\Delta = 0, \delta_1 \neq 0$
3 0 una raíz real triple:	$\Delta = 0, \delta_1 = 0$
1 1 una raíz real y dos raíces complejas conjugadas:	$\Delta < 0.$

Demostración. Sin pérdida de generalidad, podemos considerar un polinomio mónico.

Si $\Delta \neq 0$, la cantidad de raíces reales se calcula con la Proposición 2.0.0.1.

Si $\Delta > 0$, se tiene que $t \equiv 3 \pmod{4}$ y por lo tanto hay tres raíces reales distintas.

Si $\Delta < 0$, se tiene que $t \equiv 1 \pmod{4}$, por lo que hay una raíz real. Las restantes dos raíces complejas deben ser conjugadas.

Si $\Delta = 0$, hay una raíz doble. Queremos distinguir las condiciones para las cuales dicha raíz es triple. Para las cuentas siguientes, consideramos el polinomio no homogéneo

$$\begin{aligned} p(x) &= (x - r_1)(x - r_2)^2 \\ &= x^3 + (-2r_2 - r_1)x^2 + (r_2^2 + 2r_1r_2)x + (-r_2^2r_1). \end{aligned}$$

2. Polinomios, discriminantes y raíces múltiples

Los coeficientes quedan expresados en función de las raíces como

$$\begin{aligned}a &= 1 \\b &= -2r_2 - r_1 \\c &= r_2^2 + 2r_1r_2 \\d &= -r_2^2r_1.\end{aligned}$$

Tras reemplazar estos valores en la definición de δ_i y simplificar, se obtienen las siguientes expresiones

$$\begin{aligned}\delta_1 &= -(r_2 - r_1)^2 \\ \delta_2 &= 2r_2(r_2 - r_1)^2 \\ \delta_3 &= -r_2^2(r_2 - r_1)^2.\end{aligned}$$

Comprobamos entonces que si $\delta_1 = 0$, se tiene una raíz triple. En cambio, si $\delta_1 \neq 0$, la raíz doble no coincide con la raíz simple. \square

Podemos dar una prueba alternativa del resultado aplicando el Teorema de Sturm A.0.0.1 cuyo enunciado y definiciones previas pueden leerse en el apéndice A.

Demostración. Consideramos un polinomio mónico en una variable $p(x) = x^3 + bx^2 + cx + d$. La sucesión de Sturm es

$$\begin{aligned}f_0(x) &= x^3 + bx^2 + cx + d \\ f_1(x) &= 3x^2 + 2bx + c \\ f_2(x) &= \frac{2}{9}(b^2 - 3c)x - d + \frac{bc}{9} = \frac{2}{9}\delta_1x - d + \frac{bc}{9} \\ f_3(x) &= \frac{9\Delta}{4\delta_1^2}.\end{aligned}$$

Si $\Delta = 0$, el polinomio tiene una raíz doble. Por el grado, solo puede ser una raíz real. Además, dicha raíz doble puede ser triple. Distinguiremos estos casos según si δ_1 se anula o no.

Si $\delta_1 = 0$, podemos despejar

$$\delta_1 = 0 \implies 3c - b^2 = 0 \implies c = \frac{b^2}{3}$$

y reescribir

$$\begin{aligned}f_2(x) &= \frac{b^3}{27} - d \text{ y} \\ \Delta &= -\frac{(b^3 - 27d)^2}{27}.\end{aligned}\tag{2.1.0.6}$$

La condición $\Delta = 0$ implica que $f_2(x) = 0$. Por la definición de los términos de la sucesión de Sturm, esto significa que p' divide a p . De esta manera, todas las raíces de p' son raíces de p . Calculamos el discriminante de p' ,

$$\Delta_2 = 4b^2 - 12c = -4\delta_1.$$

Como p' es un polinomio de grado 2, la condición $\delta_1 = 0$ implica que su discriminante se anula y que por lo tanto tiene una raíz real doble. Concluimos que p tiene una raíz real triple.

Si $\delta_1 \neq 0$, el discriminante de p' no se anula y deducimos que tiene raíces simples. Por el grado del polinomio, deducimos que p tiene todas sus raíces reales, una doble y otra simple.

Si $\Delta < 0$, el cálculo de la sucesión de Sturm finaliza en $f_2(x) = \frac{bc}{9}$. Distinguiremos dos casos según si δ_1 se anula o no.

Si $\delta_1 = 0$, analizamos la variación de signos en la siguiente tabla

	f_0	f_1	f_2	
			$f_2(x) > 0$	$f_2(x) < 0$
$\lim_{x \rightarrow -\infty} f_i(x)$	-	+	+	-
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+	-

En el primer caso, la variación en $-\infty$ es $V_{-\infty} = 1$ y la variación en $+\infty$ es $V_{+\infty} = 0$. Por el teorema de Sturm, la cantidad de raíces reales es $V_{-\infty} - V_{+\infty} = 1$. Por el grado del polinomio, las dos raíces restantes deben ser complejas conjugadas. En el segundo, $V_{-\infty} = 2$ y $V_{+\infty} = 1$, por lo que deducimos que el polinomio tiene una única raíz real y obtenemos el mismo *rrmp*.

Cuando $\delta_1 \neq 0$, distinguimos dos casos según su signo. Lo analizamos en la siguiente tabla.

	f_0	f_1	f_2		f_3
			$\delta_1 > 0$	$\delta_1 < 0$	
$\lim_{x \rightarrow -\infty} f_i(x)$	-	+	-	+	-
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+	-	-

En ambos casos, $V_{-\infty} = 2$ y $V_{+\infty} = 1$ y deducimos el mismo *rrmp*.

Si $\Delta > 0$, resulta que $\delta_1 \neq 0$. De no ser así, el desarrollo (2.1.0.6) arroja un discriminante negativo y contradice la suposición inicial. Más aún, debe ser $\delta_1 > 0$.

Supongamos $\delta_1 < 0$. Se puede pensar al discriminante como un polinomio de segundo grado en la variable d

$$\Delta(d) = -27d^2 + (18bc - 4b^3)d + c^2(b^2 - 4c).$$

Al calcular su determinante $\tilde{\Delta} = 16(b^2 - 3c)^3$, observamos que resulta negativo bajo la hipótesis. Por lo que el polinomio $\Delta(d)$ no tiene raíces reales y su signo resulta constante.

2. Polinomios, discriminantes y raíces múltiples

Para determinarlo, evaluamos en $d = 0$

$$\Delta(0) = b^2 - 4c.$$

Como $b^2 - 3c < 0$, se deduce que $c > 0$ y, por lo tanto, $b^2 - 4c < 0$. Esto contradice la suposición inicial de que el discriminante es positivo.

Como $\Delta > 0$, solo puede ser $\delta_1 > 0$ y analizamos la variación de signos de la sucesión de Sturm en la siguiente tabla.

	f_0	f_1	f_2	f_3
$\lim_{x \rightarrow -\infty} f_i(x)$	-	+	-	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+	+

Como $V_{-\infty} = 3$ y $V_{+\infty} = 0$, por el teorema de Sturm, se deduce que hay tres raíces reales distintas. \square

Ejemplo 2.1.0.1. Sea el polinomio $p_1(x, y) = x^3 + 2x^2y - xy^2 - 2y^3$, su discriminante vale $36 > 0$. Por la proposición anterior, el polinomio tiene tres raíces reales distintas. Su *rrmp* es $(111|0)$.

Para el polinomio $p_2(x, y) = x^3 - x^2y + xy^2 - 1y^3$, el discriminante es $-16 < 0$ por lo que concluimos que tiene una raíz real y una raíz compleja con su conjugada. Su *rrmp* es $(1|1)$.

Sean ahora $p_3(x) = x^3 - 3xy^2 - 2y^3$ y $p_4(x) = x^3 + 3x^2y + 3xy^2 + y^3$. El discriminante de ambos polinomios se anula por lo que tienen *rrmp* $(12|0)$. Sin embargo, si calculamos el valor de δ_1 para p_4 , obtenemos que se anula

$$\delta_1 = 3 \cdot 1 \cdot 3 - 3^2 = 0.$$

El polinomio tiene en realidad una única raíz triple y su *rrmp* es $(3|0)$.

2.2. Polinomios de grado cuatro

Según [16], un polinomio $f_n = x^n + a_1x^{n-1}y + a_2x^{n-2}y^2 + \dots + a_ny^n$ puede traducirse a un polinomio $g_n = f(x - \frac{a_1}{n}, y) = x^n + b_2x^{n-2}y^2 + b_3x^{n-3}y^3 + \dots + b_ny^n$. El *rrmp* de ambos polinomios coincide.

Definición 2.2.0.1. Dado $p(x, y) = x^4 + bx^2y^2 + cxy^3 + dy^4$ con coeficientes reales, se calcula el discriminante

$$\Delta = 16b^4d - 4b^3c^2 - 128b^2d^2 + 144bc^2d - 27c^4 + 256d^3.$$

Se define además la cantidad

$$\delta = 8bd - 9c^2 - 2b^3.$$

Proposición 2.2.0.1. *Se distinguen los siguientes rrmp para un polinomio de grado cuatro:*

1111 0	4 raíces reales distintas:	$\Delta > 0, \delta > 0, b < 0,$
112 0	1 raíz real doble y 2 raíces reales distintas:	$\Delta = 0, \delta > 0, b < 0,$
22 0	2 raíces reales dobles:	$\Delta = 0, \delta = 0, b < 0, c = 0,$
13 0	1 raíz real simple y 1 raíz real triple:	$\Delta = 0, \delta = 0, b < 0, c \neq 0,$
4 0	1 raíz real cuádruple:	$\Delta = 0, \delta = 0, b = 0,$
11 1	2 raíces reales distintas y 2 raíces complejas conjugadas:	$\Delta < 0,$
2 1	1 raíz real doble y 2 raíces complejas conjugadas:	$\Delta = 0, \delta < 0,$
0 2	2 raíces complejas conjugadas dobles:	$\Delta = 0, \delta = 0, b > 0,$
0 11	4 raíces complejas distintas:	$\Delta > 0, (\delta \leq 0 \text{ o } b > 0).$

Demostración. Sea $p(x) = x^4 + bx^2 + cx + d$ un polinomio mónico no homogéneo.

Si $\Delta < 0$, la Proposición 2.0.0.1 arroja que la cantidad de raíces reales t cumple que $t \equiv 2 \pmod{4}$. Como $\Delta \neq 0$, solo puede haber dos raíces reales distintas simples. Por el grado, las restantes dos raíces son complejas conjugadas. Esto nos permite obviar el caso en el que el discriminante es negativo en el análisis posterior.

Computaremos la sucesión de Sturm en dos casos distintos según si el coeficiente b es nulo o no.

Caso $b = 0$.

Obtenemos además formas simplificadas para las cantidades

$$\Delta = -27c^4 + 256d^3 \text{ y } \delta = -9c^2.$$

En este caso, δ no puede ser positivo.

Distinguiremos los posibles rrmp según si los coeficientes c, δ y Δ se anulan y sus posibles signos.

Si $c = 0$, tenemos que $\delta = 0$. El polinomio estudiado se escribe como $p(x) = x^4 + d$.

Si $\Delta = 256d^3 > 0$, entonces $d > 0$ y el polinomio tiene dos pares de raíces complejas conjugadas.

Si $\Delta = 256d^3 = 0$, entonces $d = 0$ y el polinomio tiene una única raíz real con multiplicidad 4.

2. Polinomios, discriminantes y raíces múltiples

Si $c \neq 0$ y $\Delta \neq 0$, la sucesión de Sturm resulta

$$f_0(x) = x^4 + cx + d$$

$$f_1(x) = 4x^3 + c$$

$$f_2(x) = -3cx + d$$

$$f_3(x) = \frac{\Delta}{27c^3}.$$

Basta con estudiar las condiciones $\Delta > 0$ y $\delta < 0$. Disponemos el análisis de variación de signos en la siguiente tabla.

	f_0	f_1	f_2		f_3	
			$c > 0$	$c < 0$	$c > 0$	$c < 0$
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	+	-	+	-
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	-	+	+	-

En el caso en que el coeficiente c es negativo, la cantidad de veces que varía el signo en $-\infty$ es 2, por lo que deducimos que $V_{-\infty} = 2$ y en $+\infty$ es 2, por lo que $V_{+\infty} = 2$. Según el teorema de Sturm, la cantidad de raíces reales distintas es $V_{+\infty} - V_{-\infty} = 0$. Como $\Delta \neq 0$, no puede haber raíces dobles. Concluimos que el polinomio tiene dos pares de raíces complejas conjugadas.

Cuando el coeficiente c es positivo, $V_{-\infty} = V_{+\infty} = 1$ por lo que deducimos que el polinomio no tiene raíces reales. De manera análoga al caso anterior, concluimos que el *rrmp* es 0[11].

Si $\Delta = 0$, la sucesión de Sturm solo tiene los siguientes tres términos

$$\begin{aligned} f_0(x) &= x^4 + cx + d \\ f_1(x) &= 4x^3 + c \\ f_2(x) &= -3cx + d \end{aligned}$$

Disponemos el análisis de variación de signos según el signo de c en la siguiente tabla.

	f_0	f_1	f_2	
			$c > 0$	$c < 0$
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	+	-
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	-	+

Cuando el coeficiente c es positivo, $V_{-\infty} = 2$ y $V_{+\infty} = 1$. Cuando el coeficiente c es negativo, $V_{-\infty} = 1$ y $V_{+\infty} = 0$. En ambos casos, concluimos que hay una única raíz real. Como el discriminante es nulo, puede tratarse de una raíz doble o cuádruple. Para distinguirlo, analizamos las raíces de su derivada p' . Como $c \neq 0$, $p'(x) = f_1(x) = 4x^3 + c$ tiene una única raíz real. Deducimos entonces que el polinomio tiene una raíz real doble y un par de raíces complejas conjugadas.

Caso $b \neq 0$.

Los términos de la sucesión de Sturm son

$$\begin{aligned} f_0(x) &= x^4 + bx^2 + cx + d = p(x) \\ f_1(x) &= 4x^3 + 2bx + c = p'(x) \\ f_2(x) &= -\frac{1}{2}bx^2 - \frac{3}{4}cx - d \\ f_3(x) &= \frac{8bd - 9c^2 - 2b^3}{b^2}x + \frac{c(b^2 + 12d)}{b^2} = \frac{\delta}{b^2}x - \frac{c(b^2 + 12d)}{b^2} \\ f_4(x) &= \frac{b^2\Delta}{4\delta^2}. \end{aligned}$$

Caso $\Delta > 0$

Si el discriminante Δ es positivo, la Proposición 2.0.0.1 indica que la cantidad de raíces reales t cumple que $t \equiv 0 \pmod{4}$. Puede entonces haber cuatro raíces reales distintas o ninguna raíz real. En tal caso, deberán ser distintas raíces complejas simples. Distinguiremos estos casos según el signo de δ y de b .

Si $\delta > 0$ y $b < 0$, disponemos el análisis de variación de signos en la siguiente tabla.

Como $V_{-\infty} = 4$ y $V_{+\infty} = 0$, comprobamos que hay cuatro raíces reales distintas.

A continuación, analizamos los casos correspondientes a la condición $\delta \leq 0$ o $b > 0$.

2. Polinomios, discriminantes y raíces múltiples

	f_0	f_1	f_2	f_3	f_4
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	+	-	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+	+	+

Si se cumple que $b > 0$, disponemos el análisis de variación de signos según el signo de δ en la siguiente tabla.

	f_0	f_1	f_2	f_3		f_4
				$\delta > 0$	$\delta < 0$	
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	-	-	+	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	-	+	-	+

En ambos casos, $V_{-\infty} = V_{+\infty} = 2$ y concluimos que no hay raíces reales. Como $\Delta \neq 0$, hay dos pares de raíces complejas conjugadas.

Si $\delta = 0$, el cuarto término de la sucesión de Sturm queda modificado como

$$f_3(x) = \frac{c(b^2 + 12d)}{b^2}.$$

Analizamos la variación de signos en la siguiente tabla, con la distinción de casos según el signo de f_3 .

	f_0	f_1	f_2	f_3		f_4
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	-	+	-	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	-	+	-	+

En ambos casos, $V_{-\infty} = V_{+\infty} = 2$ y concluimos que no hay raíces reales. Como $\Delta \neq 0$, hay dos pares de raíces complejas conjugadas.

Si se cumple que $b < 0$, resulta que δ solo puede ser negativo. Demostraremos a continuación que las condiciones $b < 0$ y $\delta = 0$ implican que el discriminante debe ser negativo y esto contradice la condición impuesta al inicio.

De la condición $\delta = 0$, tenemos que

$$8bd - 9c^2 - 2b^3 = 0 \implies b(8d - 2b^2) = 9c^2. \quad (2.2.0.1)$$

Como asumimos que $b < 0$, concluimos que $8d - 2b^2 < 0 \implies d < \frac{b^2}{4}$.

Podemos además despejar $c^2 = \frac{b(8d-2b^2)}{9}$ y reemplazarlo en la escritura del discriminante

$$\Delta(d) = \frac{4}{9}(576d^3 - 48b^2d^2 - 20b^4d - b^6)$$

que interpretamos como un polinomio de grado 3 en la variable d . Dicho polinomio tiene dos raíces reales que dependen del valor de b . Las denotamos

$$d_1 = -\frac{b^2}{12}, \quad d_2 = \frac{b^2}{4}.$$

En las condiciones estudiadas, $d < \frac{b^2}{4}$. Al ser $\Delta(d)$ una función continua en d , evaluaremos en valores intermedios de los intervalos $(-\infty; -\frac{b^2}{12})$ y $(-\frac{b^2}{12}; \frac{b^2}{4})$ para estudiar su signo. A partir de las siguientes evaluaciones

$$\Delta\left(-\frac{b^2}{11}\right) = -\frac{15}{1331}b^6 < 0, \quad \Delta\left(\frac{b^2}{5}\right) = -\frac{289}{125}b^6 < 0,$$

deducimos que $\Delta(d) < 0$ si $d < \frac{b^2}{4}$. Por lo tanto, las condiciones $\delta = 0$ y $b < 0$ implican que el discriminante es negativo y podemos descartar este caso.

Disponemos el análisis de variación de signos en el caso que $\delta < 0$. en la siguiente tabla.

	f_0	f_1	f_2	f_3	f_4
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	+	+	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+	-	+

Como $V_{-\infty} = V_{+\infty} = 2$, concluimos que no hay raíces reales. Como $\Delta \neq 0$, hay dos pares de raíces complejas conjugadas.

2. Polinomios, discriminantes y raíces múltiples

Caso $\Delta = 0$

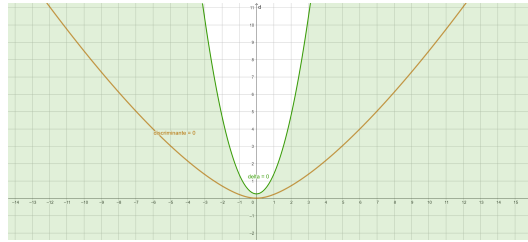
Si $\Delta = 0$, el último término de la sucesión de Sturm es nulo. En este caso, hay al menos una raíz múltiple. Si $\delta \neq 0$, hay una única raíz real doble. Se distinguen dos casos según el signo de δ .

Si $\delta > 0$ y $b < 0$, detallamos los signos de la sucesión de Sturm en la siguiente tabla.

	f_0	f_1	f_2	f_3
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	+	-
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+	+

Como $V_{-\infty} = 3$ y $V_{+\infty} = 0$, concluimos que hay tres raíces reales distintas de las cuales una es doble.

Según cálculos realizados computacionalmente, las condiciones $\Delta = 0$, $\delta > 0$ y $b > 0$ tienen intersección vacía. Lo ilustramos en el caso $b = 1$. En la figura siguiente, se realiza el gráfico implícito de las curvas dadas por $\Delta = 0$ en naranja y $\delta = 0$ en verde. Cuando evaluamos $c = 10$ y $d = 0$, $\delta = -78$ es negativo. Esto nos permite concluir que



toda la región verde está descrita por la inecuación $\delta < 0$. Por lo tanto, en el caso en que $b = 1$ no existen valores para los coeficientes c y d que cumplan en simultáneo que el discriminante se anule y que δ sea positivo.

Si $\delta < 0$, disponemos los signos de la sucesión de Sturm en la siguiente tabla.

	f_0	f_1	f_2		f_3
			$b > 0$	$b < 0$	
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	-	+	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	-	+	-

Vemos que, sin importar el signo de b , hay una única raíz real. Por el grado, esta debe ser la raíz doble y las raíces restantes son complejas conjugadas.

Si $\delta = 0$, podemos reescribir el último término de la sucesión como $f_3(x) = \frac{c(b^2+12d)}{b^2}$. Más aún, probaremos que bajo las condiciones $\delta = 0$ y $\Delta = 0$, este término es nulo. Las siguientes computaciones fueron realizadas con el software Maple [18]. La base de

Groebner del ideal generado por δ y Δ es $\{-\delta, -\Delta\}$. Comprobamos que $c(b^2 + 12d)$ no está en dicho ideal pero que $c^2(b^2 + 12d)^2$ sí lo está:

$$c^2(b^2 + 12d)^2 = \left(\frac{bc^2}{2} + 16\right)(-\delta) + \left(\frac{-b}{2}\right)(-\Delta).$$

Deducimos entonces que

$$\delta = 0, \Delta = 0 \implies c^2(b^2 + 12d)^2 = 0 \implies c(b^2 + 12d) = 0.$$

concluimos entonces que $f_3(x) = 0$. Reescribimos la sucesión de Sturm como

$$\begin{aligned} f_0(x) &= x^4 + bx^2 + cx + d \\ f_1(x) &= 4x^3 + 2bx + c \\ f_2(x) &= -\frac{1}{2}bx^2 - \frac{3}{4}cx - d. \end{aligned}$$

Si $c \neq 0$, disponemos el análisis de variación de signos en la siguiente tabla.

	f_0	f_1	f_2	
			$b > 0$	$b < 0$
$\lim_{x \rightarrow -\infty} f_i(x)$	+	-	-	+
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	-	+

Si $b < 0$, como $V_{-\infty} = 2$ y $V_{+\infty} = 0$, hay dos raíces reales distintas. Como $f_3(x)$ es nulo, el máximo común divisor entre p y p' es $f_2(x)$ y tiene grado 2, por lo que comparten dos raíces, contadas con multiplicidad. Según la multiplicidad de las raíces de su derivada, p puede tener una raíz triple o dos raíces dobles. Para estudiar las raíces de $p'(x) = 4x^3 + 2bx + c$, aplicamos la Proposición 2.1.0.1. Calculamos su discriminante que denotamos $\tilde{\Delta} = -16(8b^3 + 27c^2)$. De la condición $\delta = 0$, despejamos

$$9c^2 = 8bd - 3b^3 \implies 27c^2 = 24bd - 6b^3$$

y lo reemplazamos en el cálculo previo del discriminante $\tilde{\Delta} = -32b(b^2 + 12d)$.

Como vimos que $c(b^2 + 12d) = 0$ y en este caso asumimos la condición $c \neq 0$, deducimos que debe ser $b^2 + 12d = 0$. Por lo tanto, el discriminante $\tilde{\Delta}$ se anula. Calculamos además la cantidad δ_1 definida en (2.1.0.2) para el polinomio p' como

$$\delta_1 = 3 \cdot 4 \cdot 2b - 0^2 = 24b$$

que bajo estas condiciones no se anula. Por lo tanto, p' tiene una raíz doble y otra simple. Concluimos entonces que p tiene una raíz real triple y, por lo tanto, otra raíz real simple.

2. Polinomios, discriminantes y raíces múltiples

Si $b > 0$, como $V_{-\infty} = V_{+\infty} = 1$, no hay raíces reales. Como $\Delta = 0$, el polinomio tiene al menos una raíz doble. Si una raíz compleja es doble, su conjugado también resulta ser una raíz doble. En este caso, el polinomio tiene dos raíces complejas conjugadas dobles.

Si $c = 0$, el polinomio se escribe como $p(x) = x^4 + bx^2 + d$ y resulta ser una bicuadrática. Podemos hacer el cambio de variables $t = x^2$ y analizar la cantidad de raíces reales del polinomio $\tilde{p}(x) = t^2 + bt + d$.

Calculamos su discriminante como $b^2 - 4d$. Recordamos que como $\delta = 0$, de la ecuación (2.2.0.1) se tiene que el producto $-2b(b^2 - 4d) = 9c^2 = 0$.

Así, el polinomio tiene una raíz real doble $x_0 = \frac{-b}{2}$.

Si $b < 0$, la ecuación $t^2 = \frac{-b}{2}$ tiene dos soluciones y resulta que p tiene dos raíces reales dobles.

Si $b > 0$, la ecuación no tiene solución y el polinomio p no tiene raíces reales. Al ser un polinomio de grado 4 cuyo discriminante se anula, han de ser un par de raíz compleja y su conjugada doble.

□

3. El espacio de funciones

En este capítulo, definiremos el espacio de funciones de una red lineal convolucional. Veremos que se trata de un conjunto semi-algebraico y discutiremos su dimensión.

Consideramos la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L . Como definimos en el Capítulo 1, se trata de una tupla de vectores $\mathbf{d} = (d_0, \dots, d_L)$, $\mathbf{k} = (k_1, \dots, k_L)$ y $\mathbf{s} = (s_1, \dots, s_L)$. Las funciones que componen esta red son composiciones de L convoluciones α_l con tamaño de entrada d_{l-1} y de salida d_l , tamaño de filtro k_l y salto s_l para $l = 1, \dots, L$. Estas cantidades cumplen la ecuación (1.1.0.5).

Observación 3.0.0.1. Sean W_1, \dots, W_L matrices convolucionales con tamaño de filtro k_l y salto s_l . Cada matriz $W_l \in \mathbb{R}^{d_l \times d_{l-1}}$ y las cantidades k_l, s_l, d_{l-1} y d_l están relacionadas por la ecuación (1.1.0.5). Diremos que las matrices W_1, \dots, W_L son *compatibles*.

Al variar los pesos del filtro de cada convolución, se obtienen las distintas funciones que constituyen la red. Estos pesos son los parámetros que el algoritmo buscará ajustar para que resulten en el mejor desempeño de su tarea.

Definición 3.0.0.1. Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ el *espacio de funciones* es el conjunto $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ definido como

$$\{\overline{W} \in \mathbb{R}^{d_L \times d_0} : \overline{W} = \prod_{l=L}^1 W_l, W_l \in \mathbb{R}^{d_l \times d_{l-1}} \text{ convolucional de salto } s_l \text{ y tamaño de filtro } k_l\} \quad (3.0.0.1)$$

Se trata del conjunto de todas las matrices de dimensión $d_L \times d_0$ que pueden escribirse como un producto de L matrices convolucionales W_1, \dots, W_L con tamaño de filtro k_l , de entrada d_{l-1} , de salida d_l y salto s_l fijados por los vectores \mathbf{k}, \mathbf{d} y \mathbf{s} de la arquitectura considerada.

El espacio de funciones es entonces el conjunto de todas las transformaciones lineales que pueden ser representadas por una red lineal convolucional mediante la variación de los parámetros.

Definición 3.0.0.2. Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$, definimos el *mapa de parametrización* como la función que toma L filtros $w^l \in \mathbb{R}^{k_l}$, asociados a convoluciones α_l con salto s_l , y les asigna el producto de sus matrices en base canónica W_l

$$\begin{aligned} \mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}} : \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L} &\rightarrow \mathbb{R}^{d_L \times d_0} \\ (w^1, \dots, w^L) &\mapsto \overline{W} = W_L \dots W_1. \end{aligned}$$

3. El espacio de funciones

Observación 3.0.0.2. Fijada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$, el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es la imagen del mapa de parametrización $\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$

$$\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} = \text{Im}(\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}).$$

Como indica la Observación 1.1.0.2, la ecuación (1.1.0.5) permite deducir el tamaño de salida a partir del tamaño de entrada, del salto y del tamaño del filtro. Por esto, el espacio de funciones puede quedar definido a partir de los vectores \mathbf{k}, \mathbf{s} , la profundidad L de la red y el tamaño de entrada o de salida.

En la sección siguiente, daremos ejemplos del espacio de funciones correspondiente a distintas arquitecturas. Comenzaremos con una definición necesaria para comprender la relación entre arquitecturas.

3.1. Arquitecturas filling

Fijada la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L , las funciones del espacio de funciones asociado $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ son composiciones de L convoluciones α_l con tamaño de entrada d_{l-1} y de salida d_l , tamaño de filtro k_l y salto s_l :

$$\begin{aligned} f : \mathbb{R}^{d_0} &\rightarrow \mathbb{R}^{d_L} \\ f(x) &= (\alpha_L \circ \cdots \circ \alpha_1)(x). \end{aligned}$$

Como vimos en la Proposición 1.1.0.1, la función f también es una convolución. Más aún, podemos calcular el tamaño de su filtro k y de su salto s como

$$k = k_1 + \sum_{l=2}^L (k_l - 1) \prod_{m=1}^{l-1} s_m, \quad (3.1.0.1)$$

$$s = \prod_{l=1}^L s_l. \quad (3.1.0.2)$$

A partir de estas cantidades, podemos considerar una nueva arquitectura $((d_0, d_L), k, s)$. El espacio de funciones $\mathcal{M}_{(d_0, d_L), k, s}$ es el conjunto de matrices convolucionales de dimensiones $d_L \times d_0$ con tamaño de filtro k y salto s .

Observación 3.1.0.1. Fijada la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L , podemos calcular los valores k y s como en (3.1.0.1). De la Proposición 1.1.0.1 sobre la composición de convoluciones, deducimos que el producto de las L matrices convolucionales W_l de salto s_l y tamaño de filtro k_l es una matriz convolucional de salto s y tamaño de filtro k . Es decir, se tiene la siguiente inclusión de conjuntos de matrices

$$\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} \subseteq \mathcal{M}_{(d_0, d_L), k, s}.$$

Ambos conjuntos están conformados por matrices convolucionales de dimensiones $d_L \times d_0$ con tamaño de filtro k y salto s . Sin embargo, los elementos de $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ deben cumplir más condiciones: deben poder ser descritos como el producto de L matrices convolucionales W_l con tamaño de entrada d_{l-1} y de salida d_l , tamaño de filtro k_l y salto s_l . Si recordamos la Definición 3.0.0.2 del mapa de parametrización, esto quiere decir que todo elemento $\overline{W} \in \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ debe ser un elemento de la imagen de $\mu_{\mathbf{d},\mathbf{k},\mathbf{s}}$.

Definición 3.1.0.1. Dada la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L , se tienen los valores k y s como en (3.1.0.1) y $\mu_{\mathbf{d},\mathbf{k},\mathbf{s}}$ el mapa de parametrización como en la Definición 3.0.0.2. Sea un filtro $\overline{w} \in \mathbb{R}^k$ al que podemos asociarle la matriz convolucional \overline{W} con filtro \overline{w} y salto s . Diremos que \overline{w} es *representable* si $\overline{W} \in \text{Im}(\mu_{\mathbf{d},\mathbf{k},\mathbf{s}}) = \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$.

Dado un elemento $\overline{W} \in \mathcal{M}_{(d_0, d_L), k, s}$, no siempre podremos hallar matrices convolucionales W^l de filtro $w_l \in \mathbb{R}^{k_l}$ y salto s_l tales que

$$W = W_L \dots W_1.$$

Es decir que no cualquier filtro $\overline{w} \in \mathbb{R}^k$ es representable. Esto dependerá de la arquitectura y motiva la siguiente definición.

Definición 3.1.0.2. Sea una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L y su espacio de funciones $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$. Se tienen los valores k y s como en (3.1.0.1). A partir de estas cantidades, se considera otra arquitectura $((d_0, d_L), k, s)$ y su espacio de funciones $\mathcal{M}_{(d_0, d_L), k, s}$. La arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ se dice *filling* si vale que

$$\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}} = \mathcal{M}_{(d_0, d_L), k, s}.$$

Es decir, que todo elemento $\overline{W} \in \mathcal{M}_{(d_0, d_L), k, s}$ puede ser descrito como el producto de L matrices convolucionales con tamaño de filtro k_l y salto s_l fijados por \mathbf{k} y \mathbf{s} .

Veamos ejemplos de arquitecturas *filling* y no *filling*.

Ejemplo 3.1.0.1. Dada la arquitectura $\mathbf{d} = (3, 2, 1)$, $\mathbf{k} = (2, 2)$ y $\mathbf{s} = (1, 1)$, el espacio de funciones correspondiente $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ es un subconjunto de $\mathcal{M}_{(3,1),3,0} = \mathbb{R}^{1 \times 3}$. Los elementos en $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ se describen como el siguiente producto de dos matrices

$$\overline{W} = W_2 W_1 = \begin{bmatrix} c & d \end{bmatrix} \begin{bmatrix} a & b & 0 \\ 0 & a & b \end{bmatrix} = \begin{bmatrix} ac & ad - bc & bd \end{bmatrix} =: \begin{bmatrix} A & B & C \end{bmatrix}.$$

Si tomamos un elemento $\begin{bmatrix} A & B & C \end{bmatrix} \in \mathcal{M}_{(3,1),3,0}$, para poder escribirlo como un elemento de $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ debe ser posible hallar $a, b, c, d \in \mathbb{R}$ tales que

$$\begin{aligned} A &= ac, \\ B &= ad - bc, \\ C &= bd, \end{aligned}$$

3. El espacio de funciones

Si identificamos los filtros (a, b) y (c, d) con los coeficientes de dos expresiones lineales $ax + b$ y $cx + d$, el filtro del producto de las matrices W_1 y W_2 coincide con los coeficientes del producto de las expresiones.

$$(ax + b)(cx + d) = acx^2 + (ad + bc)x + bd = Ax^2 + Bx + C.$$

El filtro (A, B, C) se identifica entonces con los coeficientes de una cuadrática que se escribe como el producto de estas expresiones.

Por lo tanto, para poder parametrizar un elemento de $\mathcal{M}_{(3,1),3,0}$ como uno de $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$, los pesos del filtro deben identificarse con coeficientes de expresiones cuadráticas con una o dos raíces reales. Existen elementos de $\mathbb{R}^{1 \times 3}$ cuyas entradas se identifican con coeficientes de polinomios cuadráticos sin raíces reales. Por lo tanto, la arquitectura no es *filling*.

Más aún, podemos describir al conjunto de elementos (A, B, C) representables con la condición $B^2 - 4AC \geq 0$.

Ejemplo 3.1.0.2. Dada la arquitectura $\mathbf{d} = (5, 3, 2)$, $\mathbf{k} = (3, 2)$ y $\mathbf{s} = (1, 1)$, el espacio de funciones respectivo $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ es un subconjunto de $\mathcal{M}_{(5,2),4,1} \subset \mathbb{R}^{2 \times 5}$. Un elemento de $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ se escribe como el siguiente producto de matrices

$$\overline{W} = W_2 W_1 = \begin{bmatrix} d & e & 0 \\ 0 & d & e \end{bmatrix} \begin{bmatrix} a & b & c & 0 & 0 \\ 0 & a & b & c & 0 \\ 0 & 0 & a & b & c \end{bmatrix} = \begin{bmatrix} A & B & C & D & 0 \\ 0 & A & B & C & D \end{bmatrix}$$

con $(A, B, C, D) = (ad, bd + ae, cd + be, ce)$. Para poder parametrizar un elemento arbitrario de $\mathcal{M}_{(5,2),4,1}$ con filtro (A, B, C, D) , debemos hallar $a, b, c, d, e \in \mathbb{R}$ que cumplan la igualdad descrita previamente. Como en el ejemplo anterior, podemos identificar el filtro (a, b, c) con el polinomio $ax^2 + bx + c$, al filtro (d, e) con el polinomio $dx + e$ y al filtro (A, B, C, D) con los coeficientes del polinomio $Ax^3 + Bx^2 + Cx + D$. El filtro del producto de las matrices coincide con los coeficientes del producto de las expresiones

$$(ax^2 + bx + c)(dx + e) = adx^3 + (bd + ae)x^2 + (cd + be)x + ce = Ax^3 + Bx^2 + Cx + D.$$

Los elementos representables se identifican entonces con polinomios de grado 3 que pueden escribirse como el producto de un factor cuadrático y uno lineal. Como esto siempre sucede, para cualquier elección de (A, B, C, D) se pueden hallar a, b, c, d y e que cumplan lo pedido. Deducimos entonces que $\mathcal{M}_{(5,3,2),(3,2),(1,1)} = \mathcal{M}_{(5,2),4,1}$. Se trata de una arquitectura *filling*.

Estos ejemplos ilustran una identificación que se utilizará a lo largo de todo el trabajo entre matrices convolucionales y polinomios. Las primeras definiciones y resultados se darán en el caso particular de matrices convolucionales de salto 1.

3.2. Matrices convolucionales de salto 1

A lo largo de esta sección, trabajaremos con arquitecturas tales que $\mathbf{s} = (1, \dots, 1)$. Es decir, con matrices convolucionales de salto 1.

3.2.1. Identificación polinomial

Definición 3.2.1.1. Para un valor de $k \in \mathbb{N}$, la función π toma un filtro $w = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$ e identifica los pesos con los coeficientes de un polinomio homogéneo en $\mathbb{R}[x, y]_{k-1}$ como se describe a continuación

$$\pi(w) = c_0 x^{k-1} + c_1 x^{k-2} y + \dots + c_{k-2} x y^{k-2} + c_{k-1} y^{k-1}. \quad (3.2.1.1)$$

Si quisiéramos definir una función similar en el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, que a una matriz $\overline{W} \in \mathbb{R}^{d_L \times d_0}$ de filtro $\overline{w} = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$ le asigne el polinomio $\pi(\overline{w})$, podríamos recurrir al mapa φ de la definición (1.1.0.4) de la siguiente manera

$$\bar{\pi}(\overline{W}) = \pi(\varphi(\overline{W})) = \pi(\overline{w}) = c_0 x^{k-1} + c_1 x^{k-2} y + \dots + c_{k-2} x y^{k-2} + c_{k-1} y^{k-1}.$$

Como un elemento del espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ se escribe como un producto de L matrices convolucionales $W_l \in \mathbb{R}^{d_L \times d_0}$ con filtro $w^l \in \mathbb{R}^{k_l}$, para definir correctamente la función $\bar{\pi}$, debemos comprobar que

$$\bar{\pi}(W_L \dots W_1) = \bar{\pi}(W_L) \dots \bar{\pi}(W_1) = \pi(w^L) \dots \pi(w^1).$$

En los Ejemplos 3.1.0.2 y 3.1.0.1, utilizamos esta identificación y comprobamos que el filtro del producto de matrices se identificaba con el producto de los polinomios asociados a cada matriz. Esto vale en general, como veremos en la siguiente proposición.

Proposición 3.2.1.1. Sean W_1, W_2, \dots, W_L matrices convolucionales compatibles como en 3.0.0.1 y salto 1. Luego, $\overline{W} = W_L \dots W_1$ satisface que

$$\bar{\pi}(\overline{W}) = \bar{\pi}(W_L) \dots \bar{\pi}(W_1).$$

Demostración. El resultado se demuestra por inducción en la cantidad L de matrices convolucionales.

El caso base corresponde al producto de dos matrices W_1 y W_2 con respectivos filtros w^1 de tamaño k_1 y w^2 de tamaño k_2 .

De acuerdo a lo visto en (1.1.0.8), el producto $\overline{W} = W_2 W_1$ tiene un filtro u de tamaño $k = k_1 + k_2 - 1$ con pesos dados por

$$u_m = \sum_{\substack{j \in [k_2], l \in [k_1] \\ j+l=m}} w_j^2 w_l^1.$$

Esta matriz \overline{W} se identifica con el polinomio

$$\bar{\pi}(\overline{W}) = u_0 x^{k-1} + u_1 x^{k-2} y + \dots + u_{k-1} y^{k-1}$$

3. El espacio de funciones

y a cada matriz W_i le corresponde un polinomio según

$$\begin{aligned}\bar{\pi}(W_1) &= \sum_{l=0}^{k_1-1} w_l^1 x^{(k_1-1)-l} y^l \\ \bar{\pi}(W_2) &= \sum_{j=0}^{k_2-1} w_j^2 x^{(k_2-1)-j} y^j.\end{aligned}$$

Calculamos su producto

$$\begin{aligned}\bar{\pi}(W_2)\bar{\pi}(W_1) &= \sum_{j=0}^{k_2-1} \left(\sum_{l=0}^{k_1-1} w_j^2 w_l^1 x^{(k_2-1)+(k_1-1)-(j+l)} y^{j+l} \right) \\ &= \sum_{m=0}^{k-1} \left(\sum_{\substack{j \in [k_2], \\ l \in [k_1], \\ j+l=m}} w_j^2 w_l^1 \right) x^{k-1-m} y^m\end{aligned}$$

y comprobamos que coincide con $\bar{\pi}(\bar{W})$.

Consideramos el producto de L matrices convolucionales como en el enunciado, bajo la hipótesis que $\bar{\pi}(W_{L-1} \dots W_1) = \bar{\pi}(W_{L-1}) \dots \bar{\pi}(W_1)$. Notamos $\tilde{W} = W_{L-1} \dots W_1$ y por la Proposición 1.1.0.1 se trata de una matriz convolucional de salto $s = 1^L = 1$. Por lo visto en el caso base,

$$\begin{aligned}\bar{\pi}(W_L \dots W_1) &= \bar{\pi}(W_L \tilde{W}) = \bar{\pi}(W_L) \bar{\pi}(\tilde{W}) \\ &= \bar{\pi}(W_L) \bar{\pi}(W_{L-1}) \dots \bar{\pi}(W_1).\end{aligned}$$

□

Como consecuencia de la Proposición 3.2.1.1, podemos definir correctamente la función $\bar{\pi}$.

Definición 3.2.1.2. Si $\bar{W} \in \mathbb{R}^{d_L \times d_0}$ es una matriz convolucional de filtro $\bar{w} = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$, se define la función $\bar{\pi}$ como

$$\bar{\pi}(\bar{W}) = \pi(\varphi(\bar{W})) = c_0 x^{k-1} + c_1 x^{k-2} y + \dots + c_{k-2} x y^{k-2} + c_{k-1} y^{k-1}. \quad (3.2.1.2)$$

Observación 3.2.1.1. Sea $\bar{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ una matriz convolucional de filtro $\bar{w} \in \mathbb{R}^k$. Como consecuencia de las Definiciones (3.2.1.2) y 3.2.1.2 y de la Proposición 3.2.1.1, se tiene el siguiente diagrama conmutativo:

$$\begin{array}{ccc} \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} & \xrightarrow{\varphi} & \mathbb{R}^k \\ \downarrow \bar{\pi} & \swarrow \pi & \\ \mathbb{R}[x, y]_{k-1} & & \end{array}$$

Recordamos que como $\overline{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, se escribe como el producto $\overline{W} = W_L \dots W_1$ de L matrices convolucionales W_l de filtros $w^l \in \mathbb{R}^{k_l}$, $l = 1, \dots, L$.

Es decir que $\overline{W} = \mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}((w^1, \dots, w^L))$ con $\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ el mapa de parametrización definido en 3.0.0.2. Se puede definir una función sobre el espacio de parámetros $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$, $\Pi : \mathcal{W} \rightarrow \mathbb{R}[x, y]$ que tome un vector de parámetros $\theta = (w^1, \dots, w^L)$ y le asigne el polinomio $\bar{\pi}(\mu(\theta))$. Se tiene entonces el siguiente diagrama conmutativo

$$\begin{array}{ccccc} \mathcal{W} & \xrightarrow{\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}} & \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} & \xrightarrow{\varphi} & \mathbb{R}^k \\ & \searrow \Pi & \downarrow \bar{\pi} & \swarrow \pi & \\ & & \mathbb{R}[x, y]_{k-1} & & \end{array}$$

Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$, podemos identificar a las funciones representadas por la red con polinomios homogéneos que se escriben como un producto de factores específicos. La cantidad de factores depende de la profundidad de la red y el grado de cada factor depende del tamaño de los filtros indicado en el vector \mathbf{k} .

Para una arquitectura en que todas las convoluciones tienen salto 1, el siguiente resultado identifica al espacio de funciones con un conjunto de polinomios particular.

Corolario 3.2.1.1. Sean $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ una arquitectura de profundidad L con $\mathbf{s} = (1, \dots, 1)$ y $k = \sum_{i=1}^L (k_i - 1) + 1$ el tamaño del filtro de la composición resultante. Luego, $\bar{\pi}$ identifica a $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ con el conjunto de polinomios de grado k que pueden expresarse como el producto de polinomios de grados $k_1 - 1, \dots, k_L - 1$:

$$\pi(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}) = \{p = q_L \dots q_1 : q_i \in \mathbb{R}[x, y]_{k_i-1}\}.$$

Ejemplo 3.2.1.1. Dada la arquitectura determinada por $\mathbf{d} = (7, 5, 4, 3)$, $\mathbf{k} = (3, 2, 2)$ y $\mathbf{s} = (1, 1, 1)$, su espacio de funciones \mathcal{M}_1 está conformado por elementos de la forma

$$\overline{W} = \begin{bmatrix} f & g & 0 & 0 \\ 0 & f & g & 0 \\ 0 & 0 & f & g \end{bmatrix} \begin{bmatrix} d & e & 0 & 0 & 0 \\ 0 & d & e & 0 & 0 \\ 0 & 0 & d & e & 0 \\ 0 & 0 & 0 & d & e \end{bmatrix} \begin{bmatrix} a & b & c & 0 & 0 & 0 & 0 \\ 0 & a & b & c & 0 & 0 & 0 \\ 0 & 0 & a & b & c & 0 & 0 \\ 0 & 0 & 0 & a & b & c & 0 \\ 0 & 0 & 0 & 0 & a & b & c \end{bmatrix}$$

con $a, b, c, d, e, f, g \in \mathbb{R}$.

Como cada elemento tiene un filtro de tamaño 3 y dos filtros de tamaño 2, se identifica con un polinomio de grado $k = (3 - 1) + (2 - 1) + (2 - 1) + 1 = 5$ que puede expresarse como el producto de un factor cuadrático y dos factores lineales:

$$(fx + g)(dx + e)(ax^2 + bx + c)$$

3. El espacio de funciones

Dada la arquitectura determinada por $\mathbf{d} = (6, 5, 3, 2)$, $\mathbf{k} = (2, 3, 2)$ y $\mathbf{s} = (1, 1, 1)$, su espacio de funciones \mathcal{M}_2 está conformado por elementos de la forma

$$\overline{W}' = \begin{bmatrix} f & g & 0 \\ 0 & f & g \end{bmatrix} \begin{bmatrix} c & d & e & 0 & 0 \\ 0 & c & d & e & 0 \\ 0 & 0 & c & d & e \end{bmatrix} \begin{bmatrix} a & b & 0 & 0 & 0 & 0 \\ 0 & a & b & 0 & 0 & 0 \\ 0 & 0 & a & b & 0 & 0 \\ 0 & 0 & 0 & a & b & 0 \\ 0 & 0 & 0 & 0 & a & b \end{bmatrix}$$

con $a, b, c, d, e, f, g \in \mathbb{R}$.

En este espacio, los elementos también tienen un filtro de tamaño 3 y dos filtros de tamaño 2. Esto da lugar a que cada elemento se identifique también con un polinomio de grado $k = (2 - 1) + (3 - 1) + (2 - 1) + 1 = 5$ que pueda expresarse como el producto de un factor cuadrático y dos factores lineales

$$(fx + g)(cx^2 + dx + e)(ax + b).$$

Como el producto de polinomios es conmutativo, ambos espacios se identifican con el mismo conjunto de polinomios. Esto quiere decir que dado un elemento del espacio de funciones \mathcal{M}_1 , podemos hallar filtros y las respectivas matrices convolucionales que permitan describirlo como a un elemento de \mathcal{M}_2 . Entonces, dan lugar al mismo espacio de funciones.

Observación 3.2.1.2. Consideremos dos arquitecturas $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ y $(\mathbf{d}', \mathbf{k}', \mathbf{s})$ de profundidad L con $\mathbf{s} = (1, \dots, 1)$. Si \mathbf{k} y \mathbf{k}' son dos vectores con las mismas coordenadas pero en otro orden, los elementos de cada espacio de funciones se identifican con polinomios que se escriben como un producto con la misma cantidad de factores del mismo grado. Como en el ejemplo anterior, resulta entonces que ambos espacios de funciones se identifican con el mismo conjunto de polinomios. Por lo tanto, dan lugar al mismo espacio de funciones \mathcal{M} . Lo que difiere es la forma de parametrizarlo.

A partir de esta observación, fijada una profundidad L y en el contexto de convoluciones de salto 1, podemos dar un vector $\mathbf{k} = (k_1, \dots, k_L)$ para describir el espacio de funciones generado por las distintas redes cuyas capas tienen tamaño de filtros k_1, \dots, k_L . Las posibles permutaciones de las capas dan lugar al mismo espacio de funciones.

3.2.2. Caracterización del espacio de funciones

A lo largo de esta subsección trabajaremos con convoluciones de salto 1 únicamente.

Si tenemos una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de profundidad L y calculamos $k = \sum_{i=1}^L (k_i - 1) + 1$, el espacio de funciones asociado $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es un subconjunto de $\mathcal{M}_{(d_0, d_L), k, 1}$.

Los siguientes resultados van a permitirnos describir a los elementos de $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ con ecuaciones y desigualdades sobre los coeficientes de un elemento genérico de $\mathcal{M}_{(d_0, d_L), k, 1}$.

Más aún, daremos condiciones sobre el tamaño de los filtros de una arquitectura para decidir si es *filling* o no.

Expresaremos los resultados sobre el espacio de funciones asociado a una red mediante la identificación polinomial que describimos en el Corolario 3.2.1.2. Un elemento de $\mathcal{M}_{(d_0, d_L), k, 1}$ con filtro $\bar{w} = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$ se corresponde con un polinomio homogéneo $p \in \mathbb{R}[x, y]_{k-1}$ como

$$\bar{\pi}(\mathcal{M}_{(d_0, d_L), k, 1}) = \{p = c_0 x^{k-1} + c_1 x^{k-2} y + \dots + c_{k-1} y^{k-1} \in \mathbb{R}[x, y]_{k-1}\}.$$

Como vimos en la subsección anterior, los elementos de $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ se identifican con polinomios homogéneos de grado $k-1$ que se escriben como un producto de L polinomios homogéneos q_1, \dots, q_L de grados $k_1 - 1, \dots, k_L - 1$

$$\bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}) = \{p = q_L \dots q_1 : q_i \in \mathbb{R}[x, y]_{k_i-1}\}.$$

En este contexto, dar condiciones sobre los coeficientes de una matriz de $\mathcal{M}_{(d_0, d_L), k, 1}$ es equivalente a dar condiciones sobre los coeficientes de un polinomio en $\mathbb{R}[x, y]_{k-1}$. Más aún, las condiciones sobre el tamaño de los filtros k_i se corresponden con condiciones sobre el grado de los factores q_1, \dots, q_L que aparecen en el producto que describe a la identificación de un elemento de $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$.

Lema 3.2.2.1. *Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de profundidad L con $\mathbf{s} = (1, \dots, 1)$, calculamos $k = \sum_{i=1}^L (k_i - 1) + 1$. Un elemento $\bar{W} \in \mathcal{M}_{(d_0, d_L), k, 1}$ se corresponde con una matriz convolucional de $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ si y solamente si el polinomio $p \in \mathbb{R}[x, y]_{k-1}$ con el que se identifica \bar{W} tiene al menos $e := |\{k_i : k_i \text{ es par}\}|$ raíces reales contadas con multiplicidad.*

Observación 3.2.2.1. Podemos describir esta condición en términos del *rrmp* del polinomio p que definimos en 2.0.0.4. Si p tiene *rrmp* $(\rho|\gamma)$ con $\rho = \rho_1, \dots, \rho_r$ y $\gamma = \gamma_1, \dots, \gamma_c$, entonces p se puede escribir como el producto $p = p_1^{\rho_1} \dots p_r^{\rho_r} q_1^{\gamma_1} \dots q_c^{\gamma_c}$ con $p_i \in \mathbb{R}[x, y]_1$, $i = 1, \dots, r$ y $q_j \in \mathbb{C}[x, y]_2$, $j = 1, \dots, c$. Luego p tiene al menos e raíces reales contadas con multiplicidad si y solamente si $\sum_{i=1}^r \rho_i \geq e$.

Demostración. Alcanza con demostrar que un polinomio p de grado $k-1$ se factoriza con polinomios de grados $k_1 - 1, \dots, k_L - 1$ si y solamente si tiene al menos e raíces reales contadas con multiplicidad.

Si $p = q_1^{k_1-1} \dots q_L^{k_L-1}$ entonces tiene e factores de grado impar. Cada polinomio de grado impar debe tener al menos una raíz real, por lo que p debe tener al menos e raíces reales.

Sea e' la cantidad de raíces reales de p contadas con multiplicidad. Los factores irreducibles de p son e' factores lineales y $\frac{k-1-e'}{2}$ factores cuadráticos. Como $e' \geq e$, podemos multiplicar estos factores para construir los polinomios q_i de grado $k_i - 1$ del

3. El espacio de funciones

enunciado.

Para cada polinomio q_i de grado impar, elegimos uno de los e' factores lineales y se lo asignamos. Esto es posible pues hay e polinomios de grado impar y hay por lo menos esta cantidad de factores lineales en la factorización de p . Luego, quedan $(e' - e) + \frac{k-1-e'}{2}$ factores irreducibles para asignar a los polinomios q_i hasta llegar al grado deseado $k_i - 1$. Esta tarea puede completarse pues tanto los grados de los factores irreducibles de p como los grados de los polinomios q_i suman $k - 1$. \square

Por lo tanto, a partir de este lema, podemos decidir si un elemento $\overline{W} \in \mathcal{M}_{(d_0, d_L), k, 1}$ puede ser parametrizado como un elemento en $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$. Es decir, podemos decidir si $\overline{W} \in \text{Im}(\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}})$. Consideramos el polinomio $p = \bar{\pi}(\overline{W})$ correspondiente por la identificación descrita en 3.2.1.2 y contamos, con multiplicidad, la cantidad de raíces reales que tiene. Contamos también la cantidad de filtros k_i de tamaño par de la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Es decir, contamos la cantidad de coordenadas pares del vector \mathbf{k} . Si la cantidad de raíces reales coincide o supera a la cantidad de filtros de tamaño par, concluimos que es posible parametrizar al elemento \overline{W} .

Ilustramos el resultado con las posibles arquitecturas cuyo espacio de funciones asociado es un subconjunto de $\mathcal{M}_{(d_0, d_L), 5, 1}$, es decir, del conjunto de matrices convolucionales cuyo tamaño de filtro es 5.

Ejemplo 3.2.2.1. Para dar ejemplos de distintas arquitecturas, consideraremos matrices convolucionales de salto 1 y elegiremos la cantidad de filtros y su tamaño para que el filtro de la composición resultante tenga tamaño 5. Así, para cada arquitectura, los elementos del espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ se corresponderán con polinomios de grado 4 con una estructura de raíces reales particular, determinada la cantidad de filtros de tamaño par:

- Si $L = 2$ y $\mathbf{k} = (3, 3)$, una función de la red tiene dos capas. Cada convolución tiene tamaño de filtro 3. Como no hay ningún filtro de tamaño par, los elementos del espacio de funciones asociado se identifican entonces con polinomios que tienen al menos cero raíces reales. Concluimos entonces que todo polinomio de grado cuatro está en $\pi(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}})$ y la arquitectura es *filling*.
- Si $L = 2$ y $\mathbf{k} = (4, 2)$, una función de la red tiene dos capas. Una convolución tiene un filtro de tamaño 4 y la otra de tamaño 2. Como hay dos filtros de tamaño par, los elementos del espacio de funciones asociado se identifican con polinomios con al menos dos raíces reales. Como un polinomio de grado cuatro no puede tener tres raíces reales, podemos describir al espacio de funciones como polinomios con dos o cuatro raíces reales contadas con multiplicidad. El complemento del espacio de funciones se corresponde entonces con los polinomios sin raíces reales.
- Si $L = 3$ y $\mathbf{k} = (3, 2, 2)$, una función de la red tiene dos capas. Una convolución tiene filtro de tamaño 3 y las otras dos de tamaño 2. Como hay dos filtros de tamaño par, esta arquitectura tiene el mismo espacio de funciones que la descrita

previamente. Sin embargo, la parametrización es diferente. En este caso, $\pi(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$ consiste en el producto de dos factores lineales y uno cuadrático mientras que en el caso anterior se trata del producto de un factor lineal y otro cúbico.

- Si $L = 4$ y $\mathbf{k} = (2, 2, 2, 2)$, una función de la red tiene cuatro capas. Las convoluciones tienen todas tamaño de filtro 4. Como hay cuatro filtros de tamaño par y los polinomios de grado cuatro tienen cuatro raíces, el espacio de funciones asociado se identifica con polinomios que tengan todas sus raíces reales.

Podemos entonces describir al espacio de funciones $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ asociado a una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ con condiciones sobre la cantidad de raíces reales que deben tener los polinomios en la identificación $\bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$. Esto nos va a permitir caracterizar al espacio de funciones como un subconjunto *semi-algebraico* de $\mathcal{M}_{(d_0, d_L), k, 1}$.

Definición 3.2.2.1. Un conjunto *semi-algebraico básico* en \mathbb{R}^k es un conjunto de la forma

$$\left\{ x \in \mathbb{R}^k : p(x) = 0 \wedge \bigwedge_{q \in \mathcal{Q}} q(x) > 0 \right\}, \text{ con } p \in \mathbb{R}[x], \mathcal{Q} \subset \mathbb{R}[x] \text{ finito.}$$

Un conjunto *semi-algebraico* es una unión finita de conjuntos semi-algebraicos básicos.

Dados $k, l \in \mathbb{N}$ y los conjuntos $A \subset \mathbb{R}^k, B \subset \mathbb{R}^l$, una función $f : A \rightarrow B$ se dice *semi-algebraica* si su gráfico $\text{Graf}(f)$ es un subconjunto semi-algebraico de \mathbb{R}^{k+l} .

Definimos la *dimensión* de un conjunto semi-algebraico S como el mayor entero d tal que existe una función semi-algebraica inyectiva $f : (0, 1)^d \rightarrow S$.

La cantidad de filtros de tamaño par de la arquitectura nos permitirá además determinar si es *filling* o no.

Probamos este resultado en el siguiente teorema que resume la caracterización del espacio de funciones.

Teorema 3.2.2.1. Dada $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ una arquitectura de profundidad L con $\mathbf{s} = (1, \dots, 1)$ y $k = \sum_{i=1}^L (k_i - 1) + 1$ y $s = 1$, se tiene que $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ es un subconjunto semi-algebraico de dimensión máxima de $\mathcal{M}_{(d_0, d_L), k, s}$. En particular, $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ solo se define por desigualdades.

Más aún, la arquitectura es *filling* si y solamente si a lo sumo uno de los filtros k_1, \dots, k_L tiene tamaño par.

Demostración. En primer lugar, consideramos el caso en el que no hay filtros de tamaño par, lo que resulta en $e = 0$. Como todos los polinomios tienen al menos cero raíces reales, el Lema 3.2.2.1 nos permite deducir que todo polinomio de grado $k - 1$ se corresponde con alguna matriz del espacio de funciones.

3. El espacio de funciones

Si hay un único filtro de tamaño par, tenemos que $e = 1$ y vemos que $k - 1$ es impar. Reescribimos la expresión de k

$$k = \sum_{i=1}^L k_i - (L - 1)$$

y analizamos según la paridad de $L - 1$. Si $L - 1$ es par, como hay un único filtro de tamaño par, podemos separar los términos de la sumatoria $\sum_{i=1}^L k_i$ en un término par y una cantidad par de términos impares. La sumatoria resulta par y la cantidad k resulta ser una resta de términos pares por lo que podemos concluir que es par. Si $L - 1$ es impar, podemos separar los términos de la sumatoria en un término par y una cantidad impar de términos impares. La sumatoria resulta impar por lo que la cantidad k es una resta de términos impares y podemos concluir nuevamente que es par.

Como cualquier polinomio de grado impar tiene al menos una raíz real, se cumple la condición del Lema 3.2.2.1 y todo polinomio de grado $k - 1$ se corresponde con alguna matriz del espacio de funciones.

Si hay dos o más filtros de tamaño par, podemos hallar un polinomio que no cumpla la condición del Lema 3.2.2.1. Si el grado $k - 1$ es par, podemos considerar un polinomio sin raíces reales. Si es impar, podemos considerar un polinomio con una única raíz real.

Por el lema, dicho polinomio no se corresponde con ninguna matriz convolucional del espacio de funciones y la arquitectura no es *filling*.

Como el conjunto $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ puede ser parametrizado por polinomios, por el Teorema de Tarski-Seidenberg (ver Capítulo 2 de [14]), se concluye que se trata de un conjunto semi-algebraico. El conjunto $\bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}})$ tiene dimensión máxima en $\mathbb{R}[x, y]_{k-1}$ pues pequeñas perturbaciones en los coeficientes de un polinomio no afectan su estructura de raíces. Lo desarrollamos en la siguiente demostración de la Proposición 3.2.2.1. \square

Ejemplo 3.2.2.2. Consideramos el vector $\mathbf{k} = (2, 2)$. Como $k = (2 - 1) + (2 - 1) + 1 = 3$ y hay dos filtros de tamaño par, el espacio de funciones asociado se identifica con el conjunto de polinomios de grado 2 con al menos dos raíces reales, contadas con multiplicidad. Como vimos en el Ejemplo 2.0.0.3, podemos describir esta condición gracias al signo del discriminante:

$$\{p = ax^2 + bxy + cy^2 \in \mathbb{R}[x, y]_2 : \Delta = b^2 - 4ac \geq 0\}.$$

La desigualdad que involucra el discriminante resulta ser una desigualdad sobre los coeficientes de un polinomio genérico de $\mathbb{R}[x, y]_2$. A su vez, podemos interpretarlo como una desigualdad sobre los coeficientes de un elemento genérico del espacio de convoluciones con tamaño de filtro $k = 3$.

Fijamos valores de d_0, d_1 y d_2 adecuados para poder visualizarlo. Si $d_0 = 4, d_1 = 3$ y $d_2 = 2$, los elementos de $\mathcal{M}_{(4,3,2),(2,2),(1,1)}$ conforman el subconjunto de $\mathcal{M}_{(4,3),3,1}$ siguiente

$$\left\{ \begin{bmatrix} a & b & c & 0 \\ 0 & a & b & c \end{bmatrix} : b^2 - 4ac \geq 0 \right\}.$$

Este subconjunto se obtiene como las soluciones a una desigualdad sobre los coeficientes de un elemento genérico de $\mathcal{M}_{(4,3),3,1}$ como indica el teorema.

Como hay más de un filtro de tamaño par, la arquitectura no es *filling*.

En el caso de que la arquitectura no es *filling*, nos interesa estudiar el borde del espacio de funciones.

Observación 3.2.2.2. Consideramos la topología euclídea en el espacio de polinomios $\mathbb{R}[x, y]_{k-1}$ al identificar los coeficientes de un polinomio de grado $k - 1$ con un vector de k coordenadas reales:

$$\begin{aligned} \mathbb{R}[x, y]_{k-1} &\rightarrow \mathbb{R}^k \\ p = c_0 x^{k-1} + c_1 x^{k-2} y + \dots + c_{k-1} y^{k-1} &\mapsto (c_0, \dots, c_{k-1}). \end{aligned}$$

Podemos considerar esta misma topología en el espacio de matrices convolucionales $\mathcal{M}_{(d_0, d_L), k, 1}$ al identificar una matriz convolucional \overline{W} con su filtro \overline{w} de tamaño k mediante el mapa φ :

$$\begin{aligned} \varphi : \mathcal{M}_{(d_0, d_L), k, 1} &\rightarrow \mathbb{R}^k \\ \overline{W} &\mapsto \overline{w}. \end{aligned}$$

De esta manera, la composición $\bar{\pi} = \pi \circ \varphi$

$$\begin{array}{ccc} \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} & \xrightarrow{\varphi} & \mathbb{R}^k \\ \downarrow \bar{\pi} & \swarrow \pi & \\ \mathbb{R}[x, y]_{k-1} & & \end{array}$$

resulta ser un homeomorfismo.

Proposición 3.2.2.1. *Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ una arquitectura con $e := |\{k_i : k_i \text{ es par}\}| \geq 2$ y $\mathbf{s} = (1, \dots, 1)$. Si consideramos la topología euclídea en $\mathcal{M}_{(d_0, d_L), k, s}$, el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es cerrado. Su borde consiste en matrices convolucionales que se corresponden con polinomios de *rrmp* $(\rho|\gamma)$ que satisfacen que*

$$\sum_{i=1}^r \rho_i \geq e \text{ y } |\{\rho_i : \rho_i \text{ es impar}\}| \leq e - 2. \quad (3.2.2.1)$$

Demostración. Para ver que el espacio $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es un subconjunto cerrado de $\mathcal{M}_{(d_0, d_L), k, s}$, tomaremos una sucesión convergente de polinomios en $\bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}})$. Por el Lema

3. El espacio de funciones

3.2.2.1, todos los polinomios de la sucesión tienen al menos e raíces reales. Demostraremos que el límite también y, por lo tanto, concluimos que es un elemento de $\bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$.

Podemos verlo sin pérdida de generalidad para polinomios en una variable. Sea la sucesión $(p_n)_{n \in \mathbb{N}} \in \bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$ que converge a p . Cada polinomio p_n tiene al menos e raíces reales y queremos ver que p también. Supongamos que p tiene $r' < e$ raíces reales. Luego, p tiene $k - 1 - r'$ raíces complejas. Las denotamos $z_1, \dots, z_s \in \mathbb{C}$ con su respectiva multiplicidad $m_i \in \mathbb{N}$. Como se trata de finitas raíces, podemos hallar un valor de $\varepsilon > 0$ tal que

$$B_\varepsilon(z_i) = \{|z - z_i| < \varepsilon\} \cap \mathbb{R} = \emptyset \quad \forall i = 1, \dots, s,$$

p no se anula en $\partial B_\varepsilon(z_i)$ y su único cero en el interior de $B_\varepsilon(z_i)$ es z_i . Entonces, por el Teorema de Rouché [17], para cada $i = 1, \dots, s$ vale que

$$\frac{1}{2\pi i} \int_{\partial B_\varepsilon(z_i)} \frac{p(t)}{p'(t)} dt = m_i \quad \forall i = 1, \dots, s.$$

Como $\partial B_\varepsilon(z_i)$ es compacto, la sucesión $(p_n)_{n \in \mathbb{N}}$ converge uniformemente a p en $\partial B_\varepsilon(z_i) \quad \forall i = 1, \dots, s$. Luego, tenemos que para cada $i = 1, \dots, s$,

$$\frac{1}{2\pi i} \int_{\partial B_\varepsilon(z_i)} \frac{p_n(t)}{p'_n(t)} dt \longrightarrow \frac{1}{2\pi i} \int_{\partial B_\varepsilon(z_i)} \frac{p(t)}{p'(t)} dt.$$

Luego, para cada $i = 1, \dots, s$, $\exists N_i \in \mathbb{N}$ tal que si $n \geq N_i$

$$\left| \frac{1}{2\pi i} \int_{\partial B_\varepsilon(z_i)} \frac{p_n(t)}{p'_n(t)} dt - \int_{\partial B_\varepsilon(z_i)} \frac{p(t)}{p'(t)} dt \right| < \frac{1}{2}.$$

Como ambas integrales son cantidades enteras, vale la igualdad. Si consideramos $N = \max\{N_1, \dots, N_s\}$, el polinomio p_N tiene en cada conjunto $B_\varepsilon(z_i)$ m_i raíces complejas no reales. Por lo tanto, tiene $m_1 + \dots + m_s = k - 1 - r'$ raíces complejas y deducimos que tiene $r' < e$ raíces reales. Esto contradice la hipótesis que tiene al menos r raíces reales y queda demostrado que p debe tener al menos r raíces reales.

El borde de $\bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$ consiste en elementos de $\bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$ que son límite de sucesiones contenidas en el complemento. Es decir, $p \in \partial \bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$ cumple que $p \in \bar{\pi}(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$ y existen $p^{(j)} \in \mathbb{R}[x, y]_{k-1} - \pi(\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}})$, $j \in \mathbb{N}$ tales que

$$\lim_{j \rightarrow +\infty} p^{(j)} = p.$$

Todos los polinomios $p, p^{(j)}$ tienen el mismo grado $k - 1 = \sum_{i=1}^L (k_i - 1)$. De tal sucesión, se puede extraer una subsucesión de polinomios $p^{(j)}$ con el mismo $rrmp$ ($\rho|\gamma$). Como están contenidos en el complemento, vale que

$$\sum_{i=1}^r \rho_i < e \implies \sum_{i=1}^r \rho_i \leq e - 1.$$

Como además $\sum_{i=1}^r \rho_i \leq k - 1$ y las cantidades e y $k - 1$ tienen la misma paridad, deducimos que

$$\sum_{i=1}^r \rho_i \leq e - 2.$$

En particular, vale que $|\{\rho_i : \rho_i \text{ es impar}\}| \leq e - 2$.

En el límite, los pares de raíces conjugadas complejas pueden volverse raíces reales dobles o raíces reales distintas pueden volverse iguales. Ninguno de estos dos casos incrementa la cantidad de raíces reales con multiplicidad impar. Por lo tanto, el polinomio p tiene $rrmp(\rho'|\gamma')$ que cumple

$$\sum_{i=1}^{r'} \rho'_i \geq e \quad \text{y} \quad |\{\rho'_i : \rho'_i \text{ es impar}\}| \leq e - 2.$$

Quedan demostradas las condiciones (3.3.2.1).

Consideramos ahora un elemento $p \in \mathbb{R}[x, y]$ que cumple las condiciones (3.3.2.1). Como su $rrmp(\rho|\gamma)$ cumple que $\sum_{i=1}^r \rho_i \geq e$, deducimos que $p \in \bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}})$. Para demostrar que es un elemento del borde, dado $\varepsilon > 0$, encontraremos un polinomio p_ε del complemento arbitrariamente cerca.

El polinomio p cumple además que $|\{\rho_i : \rho_i \text{ es impar}\}| \leq e - 2$. Escribimos $\rho_i = 2\alpha_i + \beta_i$ con cada $\beta_i \in \{0, 1\}$. Consideramos una factorización de $p = p_1^{\rho_1} \dots p_r^{\rho_r} q_1^{\gamma_1} \dots q_c^{\gamma_c}$ con $p_i \in \mathbb{R}[x, y]_1$ y $q_j \in \mathbb{R}[x, y]_2$ irreducibles y linealmente independientes dos a dos. Para cada $1 \leq i \leq r$ tal que ρ_i es impar, el factor $p_i^{\rho_i}$ se puede perturbar para obtener un polinomio con α_i pares de raíces complejas conjugadas y una raíz real si $\beta_i = 1$. Supongamos $p_i = x + \lambda y$ con $\lambda \in \mathbb{R}$, construimos

$$p_\varepsilon^i = ((x + (\lambda + \varepsilon i)y)(x + (\lambda - \varepsilon i)y))^{\alpha_i} (x + \lambda y).$$

Podemos construir $p_\varepsilon = (p_\varepsilon^1)^{\rho_1} \dots (p_\varepsilon^r)^{\rho_r} q_1^{\gamma_1} \dots q_c^{\gamma_c}$. De esta manera, perturbamos p para obtener un polinomio p_ε con exactamente $\sum \beta_i$ raíces reales. Como $\sum \beta_i = |\{\rho_i : \rho_i \text{ es impar}\}| \leq e - 2$, concluimos que $p_\varepsilon \notin \bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}})$. \square

A continuación, daremos ejemplos de arquitecturas y de la caracterización del espacio de funciones asociado, junto con su borde y su complemento. Expresaremos estos conjuntos a partir de desigualdades sobre los coeficientes del filtro de una matriz convolucional genérica. Como la identificación del espacio de funciones con un conjunto de polinomios homogéneos solo depende de la cantidad de capas y del tamaño de los filtros, obviaremos el vector \mathbf{d} con los tamaños de entrada y de salida en estos ejemplos. Consideraremos capas con filtros de tamaño $k_1 \geq 2$.

Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$, con $\mathbf{s} = (1, \dots, 1)$, calculamos el tamaño del filtro de un elemento $\bar{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ como $k = \sum_{i=1}^L (k_i - 1) + 1$. Este elemento se identifica con un polinomio homogéneo de grado $k - 1$. Trabajaremos con arquitecturas cuyo tamaño

3. El espacio de funciones

resultante de filtro sea $k = 3, 4, 5$ para poder identificar los elementos con polinomios homogéneos de grado 2, 3 y 4 respectivamente. Haremos uso de los posibles *rrmp* que calculamos en el Capítulo 2.

Ejemplo 3.2.2.3. Para que el filtro resultante tenga tamaño $k = 2$, se distinguen dos arquitecturas.

Si $L = 1$, $\mathbf{k} = (3)$. Esta arquitectura es *filling*.

Si $L = 2$, $\mathbf{k} = (2, 2)$, la cantidad de filtros de tamaño par es $e = 2$. Deducimos entonces que un elemento del espacio de funciones se identifica con un polinomio de grado 2 con al menos dos raíces reales. A partir del Ejemplo 2.0.0.3, deducimos que un elemento del espacio de funciones se identifica con un polinomio de grado 2 con *rrmp* $(\rho|\gamma)$ $(11|0)$ o $(2|0)$. Los elementos del borde se identifican con polinomios cuyo *rrmp* $(\rho|\gamma)$ cumple que $|\{\rho_i : \rho_i \text{ es impar}\}| = 0$, por lo que deben tener *rrmp* $(2|0)$. Finalmente, los elementos del complemento se identifican con polinomios cuyo *rrmp* es $(0|1)$. Describimos estos conjuntos con desigualdades sobre los coeficientes del filtro de las matrices convolucionales

$$\begin{aligned}\mathcal{M}_{\mathbf{d},(2,2),(1,1)} &= \{(a, b, c) \in \mathbb{R}^3 : b^2 - 4ac \geq 0\}, \\ \partial\mathcal{M}_{\mathbf{d},(2,2),(1,1)} &= \{(a, b, c) \in \mathbb{R}^3 : b^2 - 4ac = 0\}, \\ \mathcal{M}_{\mathbf{d},(2,2),(1,1)}^c &= \{(a, b, c) \in \mathbb{R}^3 : b^2 - 4ac \leq 0\}.\end{aligned}$$

Ejemplo 3.2.2.4. Para que el filtro resultante tenga tamaño $k = 4$, distinguimos tres arquitecturas posibles.

Si $L = 1$, $\mathbf{k} = (3)$ es *filling*.

Si $L = 2$, $\mathbf{k} = (3, 2)$ también es *filling* pues tiene un único filtro de tamaño par.

Si $L = 3$, $\mathbf{k} = (2, 2, 2)$, la cantidad de filtros de tamaño par es $e = 3$. Las matrices del espacio de funciones $\mathcal{M}_{\mathbf{d},(2,2,2),(1,1,1)}$ se identifican con polinomios con tres raíces reales. Sus posibles *rrmp* son $(111|0)$, $(21|0)$ y $(3|0)$. Los elementos del borde se identifican con polinomios cuyo *rrmp* $(\rho|\gamma)$ debe cumplir que $|\{\rho_i : \rho_i \text{ es impar}\}| \leq 1$. Luego, deben ser $(21|0)$ y $(3|0)$. Se trata de los polinomios de grado tres con al menos una raíz doble. Finalmente, el complemento del espacio de funciones se identifica con polinomios que tienen una única raíz real, es decir, cuyo *rrmp* es $(1|1)$.

A partir de la Proposición 2.1.0.1, describimos estos conjuntos con ecuaciones e inecuaciones sobre los coeficientes del filtro de una matriz convolucional genérica

$$\begin{aligned}\mathcal{M}_{\mathbf{d},(2,2,2),(1,1,1)} &= \{(a, b, c, d) \in \mathbb{R}^4 : \Delta \geq 0\}, \\ \partial\mathcal{M}_{\mathbf{d},(2,2,2),(1,1,1)} &= \{(a, b, c, d) \in \mathbb{R}^4 : \Delta = 0\} \\ \mathcal{M}_{\mathbf{d},(2,2,2),(1,1,1)}^c &= \{(a, b, c, d) \in \mathbb{R}^4 : \Delta \leq 0\}.\end{aligned}$$

L	k
1	(5)
2	(4,2) o (3,3)
3	(3,2,2)
4	(2,2,2,2)

Ejemplo 3.2.2.5. Para que el filtro resultante tenga tamaño $k = 5$, distinguiremos cinco arquitecturas posibles según el vector **k**

Las arquitecturas con vector **k**=(5) o (3,3) son *filling* pues no tiene filtros de tamaño par. Las arquitecturas restantes son *no filling* pues tienen al menos dos filtros de tamaño par.

Como (4,2) y (3,2,2) son arquitecturas con $e = 2$ filtros de tamaño par, los elementos del espacio de funciones respectivo se identifican con polinomios con al menos dos raíces reales. Los elementos del borde se identifican con polinomios cuyo *rrmp* ($\rho|\gamma$) debe cumplir que ningún ρ_i sea impar. Como se trata de polinomios de grado cuatro, no pueden tener una única raíz real. Luego, los elementos del complemento se identifican con polinomios que no tienen raíces reales.

En el caso en que **k** = (2,2,2,2), los elementos del espacio de funciones se identifican con polinomios que tienen $e = 4$ raíces reales. Los elementos del borde se identifican con polinomios cuyo *rrmp* ($\rho|\gamma$) debe cumplir que $|\{\rho_i : \rho_i \text{ es impar}\}| \leq 2$. Como se trata de polinomios de grado cuatro, los elementos del complemento se identifican con polinomios que tienen dos o ninguna raíz real.

Resumimos estos casos en la siguiente tabla

k	Espacio de funciones	Borde	Complemento
(3,2,2) o (4,2)	1111 0, 112 0, 22 0, 13 0, 4 0, 11 1, 2 1	22 0, 2 1, 4 0	0 2, 0 11
(2,2,2,2)	1111 0, 112 0, 22 0, 13 0, 4 0	112 0, 22 0, 13 0, 4 0	11 1, 2 1, 0 2, 0 11

Para la arquitectura (**d**, (2,2,2,2), (1,1,1,1)), podemos describir estos conjuntos con condiciones sobre los coeficientes del filtro de una matriz convolucional genérica a partir de la Proposición 2.2.0.1

$$\begin{aligned}
 \mathcal{M}_{\mathbf{d},(2,2,2,2),(1,1,1,1)} &= \{(a, b, c, d, e) \in \mathbb{R}^5 : \Delta \geq 0, \delta \geq 0, b \leq 0\}, \\
 \partial \mathcal{M}_{\mathbf{d},(2,2,2,2),(1,1,1,1)} &= \{(a, b, c, d, e) \in \mathbb{R}^5 : \Delta = 0, \delta \geq 0, b \leq 0\} \\
 \mathcal{M}_{\mathbf{d},(2,2,2,2),(1,1,1,1)}^c &= \{(a, b, c, d, e) \in \mathbb{R}^5 : \Delta < 0\}.
 \end{aligned}$$

3.3. Generalización a saltos más grandes

En esta sección, trabajaremos con redes con convoluciones de salto mayor que 1. Daremos una caracterización del espacio de funciones correspondiente y de su borde. Co-

3. El espacio de funciones

menzaremos con una generalización de la identificación polinomial definida en (3.2.1.2) para convoluciones de salto mayor que 1.

3.3.1. Identificación polinomial

Definición 3.3.1.1. Para un valor de $k \in \mathbb{N}$ y un valor de salto $s \geq 1$, la función toma un filtro $\bar{w} = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$

$$\pi_s : \mathbb{R}^k \rightarrow \mathbb{R}[x, y]_{(k-1)s}$$

$$\pi_s(\bar{w}) = c_0 x^{(k-1)s} + c_1 x^{(k-2)s} y^s + \dots + c_{k-2} x^s y^{(k-2)s} + c_{k-1} y^{(k-1)s}$$

e identifica sus pesos con los coeficientes de un polinomio homogéneo. Si \bar{W} es una matriz convolucional de salto s y filtro $\bar{w} = (c_0, \dots, c_{k-1}) \in \mathbb{R}^k$, definimos

$$\bar{\pi}_s(\bar{W}) = c_0 x^{(k-1)s} + c_1 x^{(k-2)s} y^s + \dots + c_{k-2} x^s y^{(k-2)s} + c_{k-1} y^{(k-1)s}.$$

Para matrices de salto $s = 1$, recuperamos la identificación (3.2.1.2) vía $\pi = \pi_1$.

Como en este caso, el producto de matrices convolucionales se identifica con el producto de los polinomios asociados a cada matriz.

Proposición 3.3.1.1. Dada la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L con $\mathbf{s} = (s_1, \dots, s_L)$ y $\mathbf{k} = (k_1, \dots, k_L)$, la matriz $\bar{W} = W_L \dots W_1$ tiene salto $s' = s_{L-1} \dots s_1$ y cumple que

$$\bar{\pi}(\bar{W}) = \bar{\pi}_{s'}(W_L) \bar{\pi}(W_{L-1} \dots W_1).$$

Demostración. Como podemos pensar al producto $W_{L-1} \dots W_1$ como una matriz convolucional de salto $s' = s_{L-1} \dots s_1$, alcanza con demostrar el resultado para dos matrices convolucionales de saltos arbitrarios.

Sean W_1 y W_2 dos matrices convoluciones de filtros w_1, w_2 , de tamaños k_1 y k_2 y de saltos s_1 y s_2 respectivamente. Vamos a demostrar la siguiente igualdad

$$\bar{\pi}(W_2 W_1) = \bar{\pi}_{s_1}(W_2) \bar{\pi}(W_1).$$

Escribimos la identificación polinomial de cada matriz como

$$\bar{\pi}(W_1) = \sum_{l=0}^{k_1-1} w_l^1 x^{(k_1-1)-l} y^l$$

$$\bar{\pi}_{s_1}(W_2) = \sum_{j=0}^{k_2-1} w_j^2 x^{(k_2-1-j)s_1} y^{js_1}.$$

Al calcular su producto, si notamos $k = (k_1 - 1) + (k_2 - 1)s_1$,

$$\bar{\pi}_{s_1}(W_2) \bar{\pi}(W_1) =$$

$$\sum_{m=0}^{k-1} \left(\sum_{\substack{j \in [k_2], \\ l \in [k_1], \\ js_1 + l = m}} w_j^2 w_l^1 \right) x^{k-m} y^m,$$

comprobamos que los coeficientes son las entradas del filtro de la matriz W_2W_1 dado por (1.1.0.8). El producto coincide entonces con la identificación polinomial de W_2W_1 dada por (3.2.1.2). \square

Ejemplo 3.3.1.1. Consideramos las matrices

$$W_1 = \begin{bmatrix} a & b & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a & b & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & a & b & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a & b & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a & b \end{bmatrix}, \quad W_2 = \begin{bmatrix} c & d & 0 & 0 & 0 \\ 0 & 0 & 0 & c & d \end{bmatrix}.$$

La matriz $W_1 \in \mathbb{R}^{5 \times 10}$ tiene salto 2 y la matriz $W_2 \in \mathbb{R}^{2 \times 5}$ tiene salto. Denotamos los respectivos filtros $w^1 = \begin{bmatrix} a & b \end{bmatrix}$ y $w^2 = \begin{bmatrix} c & d \end{bmatrix}$. Los identificamos como

$$\begin{aligned} \pi(w^1) &= ax + by, \\ \pi_2(w^2) &= cx^2 + dy^2. \end{aligned}$$

El producto $\overline{W} = W_2W_1 \in \mathbb{R}^{2 \times 10}$ tiene salto 6 y filtro $\overline{w} = \begin{bmatrix} ca & cb & ad & bd \end{bmatrix}$ de tamaño 4 que se identifica con el polinomio $\pi(\overline{w}) = cax^3 + cbx^2y + adxy^2 + bdy^3$. Comprobamos que el producto $\pi_2(w^2)\pi(w^1) = \pi(\overline{w})$,

$$(cx^2 + dy^2)(ax + by) = cax^3 + cbx^2y + adxy^2 + bdy^3.$$

De la demostración anterior, deducimos que el salto de la última matriz no tiene ningún efecto en el cómputo del filtro del producto, solo en el cálculo de su salto. Este hecho da lugar a los siguientes resultados que permiten relacionar los conjuntos de polinomios con los que se identifica el espacio de funciones de distintas arquitecturas.

Lema 3.3.1.1. *Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ con $\mathbf{d} = (d_0, \dots, d_L)$, $\mathbf{k} = (k_1, \dots, k_L)$ y $\mathbf{s} = (s_1, \dots, s_L)$, valen las afirmaciones siguientes.*

i) Si $\mathbf{s}' = (s_1, \dots, s_{L-1}, 1)$, entonces $\bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}) = \bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}'})$.

ii) Si $\mathbf{k} = (k_1, \dots, k_{L-1}, 1)$ y $\mathbf{k}' = (k_1, \dots, k_{L-1})$, $\mathbf{d}' = (d_0, \dots, d_{L-1})$, $\mathbf{s}' = (s_1, \dots, s_{L-1})$, entonces $\bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}) = \bar{\pi}(\mathcal{M}_{\mathbf{d}', \mathbf{k}', \mathbf{s}'})$.

Demostración. i) Consideramos $\overline{W} = W_L \dots W_1 \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ y W'_L una matriz convolucional de salto 1 con el mismo filtro que W_L . Llamamos $\overline{W}' = W'_L \dots W_1 \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}'}$. A partir de la Proposición 3.3.1.1, sabemos que $\bar{\pi}(\overline{W}) = \bar{\pi}_s(W_L)\bar{\pi}(W_{L-1} \dots W_1)$ con $s = s_{L-1} \dots s_1$.

Como vale que

$$\bar{\pi}(\overline{W}') = \bar{\pi}_s(W'_L)\bar{\pi}(W_{L-1} \dots W_1)$$

podemos concluir que las identificaciones $\bar{\pi}(\overline{W}) = \bar{\pi}(\overline{W}')$ coinciden.

3. El espacio de funciones

ii) Consideramos $\overline{W} = W_L \dots W_1 \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, por la Proposición 3.3.1.1

$$\pi(\overline{W}) = \bar{\pi}_s(W_L) \bar{\pi}(W_{L-1} \dots W_1).$$

Si $k_L = 1$, $\bar{\pi}_s(W_L)$ es una constante y no modifica la escritura de $\bar{\pi}(\overline{W})$ como producto de factores de grado $k_1 - 1, \dots, k_{L-1} - 1$ que caracteriza a los polinomios en $\bar{\pi}(\mathcal{M}_{\mathbf{d}', \mathbf{k}', \mathbf{s}'})$. Por lo tanto, la identificación coincide con la de $W_{L-1} \dots W_1 \in \mathcal{M}_{\mathbf{d}', \mathbf{k}', \mathbf{s}'}$. □

3.3.2. Caracterización del espacio de funciones

El Lema 3.3.1.1, permite relacionar los espacios de funciones correspondientes a distintas arquitecturas. A partir de esto, distinguiremos dos tipos de arquitecturas cuyo espacio de funciones sea un conjunto semi-algebraico de dimensión completa o menor.

A partir del primer inciso del Lema 3.3.1.1, podemos caracterizar el espacio de funciones correspondientes a arquitecturas que tienen únicamente el último salto de tamaño arbitrario. Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ con vector de saltos $\mathbf{s} = (1, \dots, 1, s_L)$, valen los siguientes resultados.

Teorema 3.3.2.1. *Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ una arquitectura de profundidad L y con vector de saltos $\mathbf{s} = (1, \dots, 1, s_L)$. Luego, $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es un subconjunto semi-algebraico de dimensión máxima de $\mathcal{M}_{(d_0, d_L), k, s}$ con $s = s_L$ y $k = \sum_{i=1}^L (k_i) - L + 1$. En particular, $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ se obtiene de $\mathcal{M}_{(d_0, d_L), k, s}$ como soluciones de inecuaciones únicamente. Más aún, la arquitectura es filling si y solamente si hay a lo sumo un tamaño de filtro k_1, \dots, k_L par.*

Proposición 3.3.2.1. *Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ una arquitectura con $e := |\{k_i : k_i \text{ es par}\}| \geq 2$ y $\mathbf{s} = (1, \dots, s_L)$. Si consideramos la topología euclídea en $\mathcal{M}_{(d_0, d_L), k, s}$, el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es cerrado. Su borde consiste en matrices convolucionales que se corresponden con polinomios de rmp $(\rho|\gamma)$ que satisfacen que*

$$\sum_{i=1}^r \rho_i \geq e \text{ y } |\{\rho_i : \rho_i \text{ es impar}\}| \leq e - 2. \quad (3.3.2.1)$$

Demostración. Ambos resultados se desprenden del primer inciso del Lema 3.3.1.1.

A partir de $\mathbf{s} = (1, \dots, 1, s_L)$, podemos construir $\mathbf{s}' = (1, \dots, 1, 1)$ y resulta que $\bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}) = \bar{\pi}(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}'})$. Como la arquitectura $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}'}$ tiene todos los saltos de tamaño 1, valen el Teorema 3.2.2.1 y la Proposición 3.2.2.1 de la subsección anterior. □

El segundo inciso del Lema 3.3.1.1 permite estudiar una arquitectura cuyo último filtro tiene tamaño $k_L = 1$ a partir de la arquitectura más pequeña formada por las $L - 1$ capas previas. Solo queda cubrir los casos en que $k_L > 1$ y alguno de los saltos de las $L - 1$ primeras capas tiene salto mayor que uno.

Proposición 3.3.2.2. Si $k_L > 1$ y $s_i > 1$ para algún $i \leq L - 1$, la arquitectura no es *filling*. En particular, el espacio de funciones es un subconjunto semi-algebraico de $\mathcal{M}_{(d_0, d_L), k, s}$ de dimensión no máxima.

Demostración. Consideremos el producto final $W_L W'$ donde $W' = W_{L-1} \dots W_1$ tiene tamaño de filtro $k' = k_1 + \sum_{l=2}^{L-1} (k_l - 1) \prod_{m=1}^{l-1} s_m$ y salto $s' = s_1 \dots s_{L-1}$.

Aún si la arquitectura de las primeras $L - 1$ capas fuera *filling*, el producto $W_L W'$ puede no ser una matriz convolucional arbitraria con filtro de tamaño $k = k' + (k_L - 1)s'$ y de salto $s = s_L s'$. De ser así, si identificamos los filtros de las matrices con sus respectivos polinomios, querría decir que todo polinomio p de grado $k - 1$ podría escribirse como el producto de un polinomio real p' de grado $k' - 1$ en las variables x, y y de un polinomio real p_L de grado $k_L - 1$ en las variables $x^{s'}, y^{s'}$.

Para mostrar que esto no siempre sucede, asumimos sin pérdida de generalidad que los polinomios son mónicos y contamos los grados de libertad restantes, es decir, la cantidad de coeficientes restantes para cada polinomio. Son $k - 1$ para p , $k' - 1$ para p' y $k_L - 1$ para p_L . Como $k_L > 1$ y $s' > 1$, se tiene que

$$(k' - 1) + (k_L - 1)s' - 1 = k - 1,$$

por lo que el producto de p' y p_L no tiene suficientes grados de libertad para expresar todos los polinomios mónicos p posibles. Este recuento demuestra que

$$\dim(\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}) \leq k' + k_L - 1 < k = \dim(\mathcal{M}_{(d_0, d_L), k, s}).$$

□

Observación 3.3.2.1. Para las arquitecturas no *filling* descritas en la proposición anterior, $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ tiene una dimensión estrictamente inferior a la de $\mathcal{M}_{(d_0, d_L), k, s}$ y por lo tanto está definida como la solución a ecuaciones e inecuaciones polinomiales. Al contrario, las arquitecturas no *filling* con $\mathbf{s} = (1, \dots, 1)$ dan lugar a espacios de funciones de dimensión máxima que son soluciones a inecuaciones polinomiales.

En el siguiente ejemplo, comparamos una arquitectura similar a la del Ejemplo 3.1.0.2. En este caso, el tamaño de los saltos da lugar a una arquitectura no *filling* esta vez.

Ejemplo 3.3.2.1. Sean $\mathbf{d} = (5, 2, 1)$, $\mathbf{k} = (3, 2)$ y $\mathbf{s} = (2, 1)$, el espacio de funciones $\mathcal{M} \in \mathbb{R}^{1 \times 5}$ consiste en las matrices

$$\overline{W} = W_2 W_1 = \begin{bmatrix} d & e \end{bmatrix} \begin{bmatrix} a & b & c & 0 & 0 \\ 0 & 0 & a & b & c \end{bmatrix} = \begin{bmatrix} ad & bd & ae + cd & be & ce \end{bmatrix}.$$

En términos de polinomios, tenemos

$$\bar{\pi}(\overline{W}) = (dx^2 + ey^2)(ax^2 + bxy + cy^2) \quad (3.3.2.2)$$

por lo que la arquitectura es *filling* si y solamente si todo polinomio de grado cuatro $Ax^4 + Bx^3y + Cx^2y^2 + Dxy^3 + Ey^4$ puede escribirse como el producto descrito previamente.

3. El espacio de funciones

Veamos que el espacio de funciones está caracterizado como

$$\mathcal{M} = \{(A, B, C, D, E) \in \mathbb{R}^5 : AD^2 + B^2E = BCD \text{ y } C^2 \geq 4AE\}. \quad (3.3.2.3)$$

Si $p \in \bar{\pi}(\mathcal{M})$, existen $(a, b, c, d, e) \in \mathbb{R}^5$ tales que $A = ad, B = bd, C = ae + cd, D = be, E = ce$. Al reemplazar estos valores, vemos que se cumplen las condiciones de (3.3.2.3):

$$\begin{aligned} AD^2 + B^2E &= adb^2e^2 + b^2d^2ce = bd(ae + cd)be = BCD, \\ C^2 - 4AE &= a^2e^2 + 2adce + c^2d^2 - 4adce = (ae - cd)^2 \geq 0. \end{aligned}$$

Si ahora tenemos $(A, B, C, D, E) \in \mathbb{R}^5$ que cumpla las condiciones de (3.3.2.3), vamos a hallar los coeficientes para escribir al polinomio $p = Ax^4 + Bx^3y + Cx^2y^2 + Dxy^3 + Ey^4$ como el producto dado en (3.3.2.2). Consideramos casos según la nulidad de los coeficientes.

Si $D = E = 0$, basta con tomar $a = A, b = B, c = C, d = 1$ y $e = 0$

$$x^2(Ax^2 + Bxy + Cy^2) = Ax^4 + Bx^3y + Cx^2y^2.$$

Si $E = 0, D \neq 0$, las condiciones de (3.3.2.3) implican que $BC = AD$. Al considerar $a = \frac{C}{D}, b = 1, c = 0, d = B$ y $e = D$, resulta

$$(Bx^2 + Dy^2)\left(\frac{C}{D}x^2 + xy\right) = \frac{BC}{D}x^4 + Bx^3y + Cx^2y^2 + Dxy^3 = p.$$

Si $E \neq 0$, podemos suponer sin pérdida de generalidad que vale $E = 1$. Fijamos $c = 1, e = 1, b = D$ y distinguimos cuatro casos para determinar los valores de a y d .

Si $D, B \neq 0$, elegimos $d = \frac{B}{D}$ y $a = \frac{AD}{B}$ y calculamos el producto

$$\left(\frac{B}{D}x^2 + y^2\right)\left(\frac{AD}{B}x^2 + Dxy + y^2\right) = Ax^4 + Bx^3y + \left(\frac{B}{D} + \frac{AD}{B}\right)x^2y^2 + Dxy^3 + y^4.$$

Concluimos que $\frac{B}{D} + \frac{AD}{B} = \frac{B^2 + AD^2}{BD} = C$ con un despeje en (3.3.2.3).

Si $D \neq 0, B = 0$, despejamos de las condiciones que $A = 0$. Al tomar $d = 0$ y $a = C$, se tiene

$$y^2(Cx^2 + Dxy + y^2) = Cx^2y^2 + Dxy^3 + y^4 = p.$$

Si $A \neq 0, D = 0$, deducimos de las condiciones que $B = 0$. Como $C^2 - 4A \geq 0$ y A no se anula, alguno de los valores $\frac{C}{2} \pm \sqrt{\frac{C^2}{4} - A}$ ha de ser no nulo. Le asignamos dicho valor a d y $a = \frac{A}{d}$. El producto resulta

$$(dx^2 + y^2)\left(\frac{A}{d} + y^2\right) = Ax^4 + \left(d + \frac{A}{d}\right)x^2y^2 + y^4.$$

3.3. Generalización a saltos más grandes

Calculamos el valor del coeficiente cuadrático

$$\frac{d^2 + A}{d} = \frac{C(\frac{C}{2} \pm \sqrt{\frac{C^2}{4} - A})}{\frac{C}{2} \pm \sqrt{\frac{C^2}{4} - A}} = C.$$

Si $A = 0$, $D = 0$, deducimos de las condiciones que $B = 0$ y al elegir $a = C$ y $d = 0$

$$y^2(Cx^2 + y^2) = Cx^2y^2 + y^4.$$

4. Optimización

Al introducir el concepto de red neuronal, comentamos que el modelo aprende los valores de los parámetros θ que resulten en el mejor rendimiento. ¿Qué significa "aprender" en este contexto? En este trabajo, tratamos algoritmos de *aprendizaje supervisado*.

Para esto, necesitamos un conjunto de datos de entrenamiento. Se trata de un conjunto finito de pares de elementos de entrada y de salida. Cada par consta de un elemento de entrada y del elemento de salida que queremos que la red le asigne. Por ejemplo, si queremos diseñar un algoritmo de clasificación de caracteres, el conjunto de datos de entrenamiento consiste en imágenes de caracteres y una etiqueta con el carácter representado en cada imagen. El objetivo es que la red halle los valores de los parámetros que permitan asignar a cada imagen del conjunto de entrenamiento el carácter indicado y que estos valores permitan además la correcta clasificación de nuevas imágenes, no comprendidas en el conjunto de entrenamiento.

Debemos entonces poder medir el error cometido en la asignación de los elementos de salida. Podemos hacerlo al nivel del espacio de funciones. A partir de un conjunto de datos de entrenamiento, definimos una *función de pérdida*. Para cada función del espacio de funciones y cada elemento de entrada del conjunto de datos de entrenamiento, esta función de pérdida compara el elemento de salida que asigna la función con respecto al indicado en los datos de entrenamiento y mide el error. Buscaremos la función que minimice este error. Podemos también hacerlo en el espacio de parámetros ya que cada elección de parámetros θ da lugar a una función del espacio de funciones. Definimos una *función objetivo* que para cada elección de parámetros θ , considere la función correspondiente y mida su error en la asignación respecto a un conjunto de datos de entrenamiento. En este caso, buscaremos la elección de parámetros que minimice el error.

En esta sección, fijamos una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L con $\mathbf{s} = (1, \dots, 1)$. El espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ está parametrizado por la función $\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, que también notaremos como μ , cuya definición recordamos a continuación

$$\begin{aligned} \mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}} : \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L} &\rightarrow \mathbb{R}^{d_0 \times d_L} \\ (w^1, \dots, w^L) &\mapsto \overline{W} = W_L \dots W_1. \end{aligned}$$

Cada matriz W_i es una matriz de convolución de salto 1 y filtro w^i . Denotaremos \overline{w} al filtro de \overline{W} , la convolución que se obtiene al realizar la composición de las capas de la red. Tiene tamaño de filtro k y de salto s que calculamos en la Proposición 1.1.0.1. Notamos $\mathcal{M}_{(d_0, d_L), k, s}$ al conjunto de matrices convolucionales de salto s y tamaño de

4. Optimización

filtro k . Podemos identificar una matriz convolucional \overline{W} con su filtro \overline{w} a través de la función

$$\begin{aligned}\varphi : \mathcal{M}_{(d_0, d_L), k, s} &\rightarrow \mathbb{R}^k \\ \overline{W} &\mapsto \overline{w}.\end{aligned}\tag{4.0.0.1}$$

De esta manera, podemos considerar

$$\begin{aligned}\varphi \circ \mu : \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L} &\rightarrow \mathbb{R}^k \\ \theta = (w^1, \dots, w^L) &\mapsto \overline{w}\end{aligned}$$

que a un vector de filtros le asigna el filtro de la convolución obtenida al realizar la composición de las capas de la red.

Definición 4.0.0.1. Dada una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$, el espacio de salida es $\mathcal{X} = \mathbb{R}^{d_0}$ y el espacio de llegada es $\mathcal{Y} = \mathbb{R}^{d_L}$. El espacio de parámetros se compone de vectores cuyas coordenadas son los filtros de tamaños k_1, \dots, k_L que notamos $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$. Consideramos la red neuronal de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$

$$\begin{aligned}f : \mathcal{W} \times \mathcal{X} &\rightarrow \mathcal{Y} \\ (\theta, x) &\mapsto y = \mu(\theta)x = \overline{W}x\end{aligned}$$

que asocia un par (θ, x) de valores de parámetros y de entrada a un valor de salida y .

Un *conjunto de datos de entrenamiento* consiste en un subconjunto finito de pares de valores de entrada y salida

$$\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}.$$

Una *función objetivo* es una función definida sobre el espacio de parámetros \mathcal{W} a partir de un conjunto de datos de entrenamiento \mathcal{D}

$$\mathcal{L}_{\mathcal{D}} : \mathcal{W} \rightarrow \mathbb{R}$$

que mide el desempeño del algoritmo.

En este trabajo, describiremos a la función objetivo con una *función de pérdida* $\ell_{\mathcal{D}}$

$$\ell_{\mathcal{D}} : \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} \rightarrow \mathbb{R}$$

definida sobre el espacio de funciones correspondiente a la red que mide la diferencia entre los valores de salida estimados y los esperados, indicados en el conjunto de datos de entrenamiento.

Podemos describir a la función objetivo como la siguiente composición entre el mapa de parametrización de la Definición 3.0.0.2 y la función de pérdida

$$\mathcal{L}_{\mathcal{D}} = \ell_{\mathcal{D}} \circ \mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}.\tag{4.0.0.2}$$

Lo ilustramos en el siguiente diagrama

$$\begin{array}{ccc}
\mathcal{W} & \xrightarrow{\mu_{\mathbf{d},\mathbf{k},\mathbf{s}}} & \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}} \\
& \searrow \mathcal{L}_{\mathcal{D}} & \swarrow \ell_{\mathcal{D}} \\
& \mathbb{R} &
\end{array}$$

Un ejemplo es la función de error cuadrático que definiremos en (4.3.1.2).

Observación 4.0.0.1. Para no recargar la notación, evitaremos explicitar el conjunto de datos de entrenamiento para las funciones objetivo \mathcal{L} y de pérdida ℓ .

Como la función ℓ está definida sobre el espacio de funciones mientras que la función \mathcal{L} está definida sobre el espacio de parámetros, buscaremos comprender cómo se relacionan los candidatos a mínimo de cada función en la siguiente sección.

4.1. Puntos críticos

Recordamos además la Definición 3.2.1.2 de π , la función que permite asociar un filtro $w = (c_0, c_1, \dots, c_{k-1}) \in \mathbb{R}^k$ con un polinomio en $\mathbb{R}[x, y]_{k-1}$

$$\pi(w) = c_0 x^{k-1} + c_1 x^{k-2} y + \dots + c_{k-2} x y^{k-2} + c_{k-1} y^{k-1}.$$

Podemos definir además esta función en el espacio de funciones $\mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ para asociar una matriz convolucional \bar{W} de filtro \bar{w} con un polinomio en $\mathbb{R}[x, y]_{k-1}$

$$\begin{array}{ccc}
\mathcal{W} & \xrightarrow{\mu_{\mathbf{d},\mathbf{k},\mathbf{s}}} & \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}} \\
\downarrow \pi & \swarrow \bar{\pi} = \pi \circ \varphi & \\
\mathbb{R}[x, y]_{k-1} & &
\end{array}$$

Si $\theta = (w^1, \dots, w^L)$ y \bar{w} es el filtro de la composición de las capas de la red, vimos en la Proposición 1.1.0.1 que $\pi(\bar{w}) = \pi(w^1) \cdots \pi(w^L)$, por lo que resulta que $\bar{\pi}$ está bien definida.

Observación 4.1.0.1. A lo largo de esta sección, identificaremos los filtros del espacio de parámetros con los polinomios correspondientes y expresaremos los resultados en términos de condiciones sobre estos polinomios. Para evitar recargar la notación, no siempre se explicitará la función π .

Un problema que surge es cómo, a partir de un polinomio p , determinar de qué parámetros θ proviene. Es decir, hallar todos los θ tales que $\bar{\pi}(\mu(\theta)) = p$. Dada una arquitectura, la función $\mu_{\mathbf{d},\mathbf{k},\mathbf{s}}$ no es inyectiva, por lo que el mismo polinomio p puede provenir de distintos parámetros.

Por ejemplo, fijamos una arquitectura con $\mathbf{k} = (3, 2)$. El polinomio $p = x^3 - 4x^2y + xy^2 + 6y^3$ puede escribirse como los siguientes productos

$$p = (x^2 - 5xy + 6y^2)(x + y) = (x^2 - xy - 2y^2)(x - 3y).$$

4. Optimización

Esto quiere decir que el filtro $(1, -4, 1, 6)$ está parametrizado por $\theta = ((1, -1, -2), (1, -3))$ y por $\tilde{\theta} = ((1, -5, 6), (1, 1))$. Más aún, cualquier

$$\theta_\kappa = (\kappa(1, -1, -2), \frac{1}{\kappa}(1, -3)), \quad \kappa \in \mathbb{R}, \kappa \neq 0$$

parametriza al filtro $(1, -4, 1, 6)$. Los filtros θ_κ dan lugar a la siguiente escritura de p como un producto

$$p = (\kappa(x^2 - xy - 2y^2))(\frac{1}{\kappa}(x - 3y))$$

cuyos factores son múltiplos de los factores que aparecen en la escritura de $\pi(\theta)$.

Este ejemplo motiva la siguiente definición.

Definición 4.1.0.1. Sea una red de profundidad L con una arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Sea k el tamaño del filtro de las convoluciones que se obtienen al componer las capas de la red, que se calcula a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Dado un filtro $\bar{w} \in \mathbb{R}^k$, se denomina *fibra* al conjunto de parámetros θ tales que $\varphi(\mu(\theta)) = \bar{w}$.

Diremos que dos parámetros $\theta = (w^1, \dots, w^L)$ y $\theta' = (v^1, \dots, v^L)$ son *equivalentes a menos de factores escalares* si existe $(\kappa_1, \dots, \kappa_L) \in (\mathbb{R}^*)^L$ con $\prod_{i=1}^L \kappa_i = 1$ tal que

$$w^i = \kappa_i v^i.$$

También diremos que θ' es un *reescalamiento* de θ .

Sean $\theta \in \mathcal{W}$ y $\mu(\theta) \in \mathcal{M}$, consideramos el gradiente de la función objetivo $d_\theta \mathcal{L} = d\mathcal{L}(\theta) \in \mathbb{R}^{(k_1 + \dots + k_L) \times 1}$, el gradiente de la función de pérdida $d_{\mu(\theta)} \ell \in \mathbb{R}^{d_0 d_L \times 1}$ y la diferencial del mapa de parametrización $d_\theta \mu \in \mathbb{R}^{d_0 d_L \times (k_1 + \dots + k_L)}$. El gradiente de la función objetivo se calcula con la regla de la cadena como

$$d_\theta \mathcal{L} = d_{\mu(\theta)} \ell \circ d_\theta \mu.$$

Los puntos críticos de la función objetivo \mathcal{L} son filtros $\theta = (w^1, \dots, w^L)$ en los que $d_\theta \mathcal{L} = 0$. Nos interesa comprender si también son puntos críticos de ℓ . Es decir, si θ es tal que $d_{\mu(\theta)} \ell = 0$. Como la diferencial $d\mu$ puede anularse en un vector no nulo, puede dar lugar a puntos críticos que dependen de la parametrización y no propiamente de la función ℓ que se desea optimizar.

Ejemplo 4.1.0.1. Consideramos la arquitectura del Ejemplo (3.1.0.2) con $\mathbf{d} = (5, 3, 2)$, $\mathbf{k} = (3, 2)$ y $\mathbf{s} = (1, 1)$. Si identificamos las matrices convolucionales con su filtro, podemos mirar el mapa $\varphi \circ \mu$

$$\begin{aligned} \varphi \circ \mu : \mathbb{R}^3 \times \mathbb{R}^5 &\rightarrow \mathbb{R}^4 \\ \theta = ((a, b, c), (d, e)) &\mapsto (ad, bd + ae, cd + be, ce). \end{aligned}$$

Para cada $\theta \in \mathbb{R}^3 \times \mathbb{R}^5$, la diferencial $d_\theta \varphi \circ \mu \in \mathbb{R}^{4 \times 5}$ es de la forma

$$d_\theta \varphi \circ \mu = \begin{bmatrix} d & 0 & 0 & a & 0 \\ e & d & 0 & b & a \\ 0 & e & d & c & b \\ 0 & 0 & e & 0 & e \end{bmatrix}.$$

En el caso de tener rango máximo, $\dim(\text{Im}(d_\theta \varphi \circ \mu)) = 4$ y, por el Teorema de la dimensión, su núcleo es trivial.

Distinguiamos entonces los puntos críticos que dependen de la parametrización en la siguiente definición.

Definición 4.1.0.2. Sea $\theta = (w_1, \dots, w_L)$ un punto crítico de \mathcal{L} , es decir, tal que $d_\theta \mathcal{L} = 0$,

- i. diremos que θ es un *punto crítico puro*, si $\mu(\theta)$ es también un punto crítico de la función de pérdida restringida al espacio de funciones, es decir, si $d_{\mu(\theta)} \ell = 0$, y
- ii. diremos que θ es un *punto crítico espurio* si $d_\theta \mu$ no tiene rango máximo.

Observación 4.1.0.2. Vamos a comprobar que alcanza con la condición de que la matriz diferencial $d_\theta \mu$ tenga rango máximo para que un punto crítico de \mathcal{L} sea puro.

Sea una red de profundidad L con arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ tal que $\mathbf{s} = (1, \dots, 1)$ y $k = \sum_{i=1}^L (k_i) - (L - 1)$. Podemos considerar el mapa

$$\varphi \circ \mu : \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L} \rightarrow \mathbb{R}^k$$

que dado un vector de filtros le asigna el filtro de la convolución resultante al componer las capas de la red.

Fijamos $\theta \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$. La matriz diferencial $d_\theta \varphi \circ \mu$ tiene $\sum_{i=1}^L k_i$ columnas y $k = \sum_{i=1}^L (k_i) - (L - 1)$ filas. Si $d_\theta \mu$ tiene rango máximo, entonces $d_\theta \varphi \circ \mu \in \mathbb{R}$ también. En este caso, $\dim(\text{Im}(d_\theta \varphi \circ \mu)) = k$ y por el Teorema de la dimensión, el núcleo es trivial. Luego,

$$d_\theta \mathcal{L} = d_{\mu(\theta)} \ell \circ d_\theta \mu = 0 \implies d_{\mu(\theta)} \ell = 0$$

y θ es un punto crítico puro.

Ejemplo 4.1.0.2. Ilustramos la existencia de puntos críticos espurios que pueden dar lugar a mínimos locales en este ejemplo. Dichos mínimos locales no son de interés pues no minimizan la función de pérdida ℓ definida sobre el espacio de funciones.

Consideremos una red de dos capas con filtros de tamaño $\mathbf{k} = (2, 3)$. Esta arquitectura es *filling* y el espacio de funciones se identifica con el conjunto de polinomios homogéneos de grado 3. Está parametrizado de manera que escribiremos estos polinomios como un producto de dos factores, uno lineal y otro cuadrático. Trabajaremos identificando

4. Optimización

filtros con polinomios, sin considerar las dimensiones de las matrices convolucionales. Un elemento del espacio de parámetros es de la forma $\theta = ((a, b), (c, d, e))$ y se identifica con el polinomio $(ax + by)(cx^2 + dxy + ey^2)$.

Fijamos un elemento u del espacio de parámetros

$$u = ((1, 1), (1, 0, \frac{1}{10})).$$

Este elemento parametriza al filtro $\bar{w} = (1, 1, \frac{1}{10}, \frac{1}{10})$ que se identifica con el polinomio $\pi(\bar{w}) = x^3 + x^2y + \frac{1}{10}xy^2 + \frac{1}{10}y^3$.

Queremos que esta red aproxime a la convolución de filtro \bar{w} . Es decir, que halle valores $((a, b), (c, d, e))$ para que el filtro de la convolución resultante de la composición de las capas sea el filtro \bar{w} . Lo denominamos *filtro objetivo*. Si identificamos los filtros con los respectivos polinomios, buscamos un polinomio lineal y uno cuadrático cuyo producto sea $\pi(\bar{w})$, que llamamos *polinomio objetivo*. Cabe observar que hay infinitas combinaciones de un factor lineal y uno cuadrático cuyo producto coincide con el polinomio objetivo. Consideramos entonces una función que mide la distancia entre $\pi(\bar{w})$ y un polinomio homogéneo, escrito como el producto de un factor lineal y otro cuadrático,

$$\ell((x + ay)(bx^2 + cxy + dy^2)) = (b - 1)^2 + (ab + c - 1)^2 + (ac + d - \frac{1}{10})^2 + (ad - \frac{1}{10})^2.$$

Esta función computa la suma del cuadrado de la diferencia entre los coeficientes de un polinomio homogéneo de grado 3 y el polinomio objetivo. Elegimos que el factor lineal sea mónico para evitar reescalamientos de los filtros. Esto no afecta al cómputo de puntos críticos. Queda entonces definida la función \mathcal{L} sobre el espacio de parámetros como

$$\begin{aligned} \mathcal{L} : \mathbb{R}^4 &\rightarrow \mathbb{R} \\ (a, b, c, d) &\mapsto (b - 1)^2 + (ab + c - 1)^2 + (ac + d - \frac{1}{10})^2 + (ad - \frac{1}{10})^2. \end{aligned}$$

Un cómputo directo indica que \mathcal{L} tiene diez puntos críticos complejos, de los cuales tres corresponden a distintas escrituras del polinomio objetivo $\pi(\bar{w}) = \pi(\varphi(u))$ como un producto. Las restantes siete escrituras corresponden a puntos críticos espurios que dan lugar a factorizaciones de siete polinomios cúbicos distintos al polinomio objetivo. Uno de ellos es

$$a \approx 0,0578445483987; \quad b \approx 1,0000187825172; \quad c \approx 0,941829719725; \quad d \approx 0,0511336556138.$$

La matriz Hessiana de \mathcal{L} es definida positiva en este punto y concluimos que es un mínimo local.

Observación 4.1.0.3. Para dar condiciones que permitan distinguir los puntos críticos puros, identificaremos los filtros con polinomios. A partir de la función π definida en

(3.2.1.2), tenemos la aplicación

$$\begin{aligned} (\pi, \dots, \pi) : \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L} &\rightarrow \mathbb{R}[x, y]_{k_1-1} \times \dots \times \mathbb{R}[x, y]_{k_L-1} \\ (w^1, \dots, w^L) &\mapsto (\pi(w^1), \dots, \pi(w^L)). \end{aligned}$$

Podemos considerar entonces el mapa de parametrización sobre estos conjuntos de polinomios

$$\begin{aligned} \bar{\mu} : \mathbb{R}[x, y]_{k_1-1} \times \dots \times \mathbb{R}[x, y]_{k_L-1} &\rightarrow \mathbb{R}[x, y]_{k-1} \\ (\pi(w^1), \dots, \pi(w^L)) &\mapsto \pi(\bar{w}) = \pi(w^1) \dots \pi(w^L). \end{aligned}$$

Lo resumimos en el siguiente diagrama

$$\begin{array}{ccc} \mathcal{W} & \xrightarrow{\mu} & \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} \\ \downarrow (\pi, \dots, \pi) & & \downarrow \bar{\pi} = \pi \circ \varphi \\ \mathbb{R}[x, y]_{k_1-1} \times \dots \times \mathbb{R}[x, y]_{k_L-1} & \xrightarrow{\bar{\mu}} & \mathbb{R}[x, y]_{k-1} \end{array}$$

De esta manera, caracterizaremos la imagen de la diferencial del mapa de parametrización $\mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ en un punto $\theta = (w^1, \dots, w^L)$ a través del estudio de la diferencial de $\bar{\mu}$ en el punto $(\pi(w^1), \dots, \pi(w^L))$

$$d\bar{\mu} : \mathbb{R}[x, y]_{k_1-1} \times \dots \times \mathbb{R}[x, y]_{k_L-1} \rightarrow \mathbb{R}[x, y]_{k-1}.$$

Describiremos su imagen en la Proposición (4.1.0.1) con condiciones sobre las raíces de los polinomios $\pi(w^1), \dots, \pi(w^L)$. Para demostrar este resultado, necesitaremos una versión homogénea de la identidad de Bézout.

Lema 4.1.0.1. Sean p_1, \dots, p_n polinomios homogéneos en $\mathbb{R}[x, y]$ de grado arbitrario, $g = \text{mcd}(p_1, \dots, p_n)$ y $l = \text{mcm}(p_1, \dots, p_n)$. Para cualquier $d \geq \deg(l) - 1$, los siguientes conjuntos de polinomios homogéneos en $\mathbb{R}[x, y]$ coinciden:

$$I_d := \{\alpha_1 p_1 + \dots + \alpha_n p_n : \alpha_i \in \mathbb{R}[x, y]_{d-\deg(p_i)}\}, \quad J_d := \{hg : h \in \mathbb{R}[x, y]_{d-\deg(g)}\}$$

Demostración. En el caso en que los polinomios g y l coinciden, $p_1 = \dots = p_n$ salvo constante, y la afirmación vale.

Supongamos $\deg(g) < \deg(l)$. Para cualquier d como en el enunciado, $I_d \subset J_d$ pues cualquier combinación algebraica de los polinomios p_1, \dots, p_n es múltiplo de g .

Vamos probar que $J_d \subset I_d$ por inducción en $d \geq \deg(l) - 1$. Para probar que $J_{\deg(l)-1} \subset I_{\deg(l)-1}$, aplicaremos la identidad de Bézout con polinomios de una variable. Podemos además "deshomogeneizar" los polinomios. Podemos asumir que la forma lineal y no es

4. Optimización

un factor del mínimo común múltiplo l . De no ser, podríamos aplicar una transformación lineal inversible de la forma

$$x' = x, \quad y' = y + tx$$

para un valor de t apropiado.

Obtenemos entonces polinomios de $\mathbb{R}[x]$ como

$$\begin{aligned}\tilde{p}_i(x) &= p_i(x, 1) \\ \tilde{g}(x) &= g(x, 1) \\ \tilde{l}(x) &= l(x, 1).\end{aligned}$$

Notamos que estos polinomios tienen el mismo grado que los polinomios homogéneos correspondientes. Además, $\tilde{g} = \gcd(\tilde{p}_1, \dots, \tilde{p}_n)$ y $\tilde{l} = \text{mcm}(\tilde{p}_1, \dots, \tilde{p}_n)$.

Por la identidad de Bézout con $\tilde{g}, \tilde{p}_1, \dots, \tilde{p}_n$, existen $\tilde{\beta}_1, \dots, \tilde{\beta}_n \in \mathbb{R}[x]$ tales que

$$\tilde{g} = \tilde{\beta}_1 \tilde{p}_1 + \dots + \tilde{\beta}_n \tilde{p}_n.$$

Luego, para cualquier $\tilde{h} \in \mathbb{R}[x]$,

$$\tilde{h}\tilde{g} = \tilde{\alpha}_1 \tilde{p}_1 + \dots + \tilde{\alpha}_n \tilde{p}_n. \quad (4.1.0.1)$$

con $\tilde{\alpha}_i \in \mathbb{R}[x]$.

Como $\deg(\tilde{h}\tilde{g}) \leq \deg(\tilde{l}) - 1$, podemos suponer que $\deg(\tilde{\alpha}_i) < \deg(\tilde{l}) - \deg(\tilde{p}_i)$. De no ser así, dividimos a cada polinomio $\tilde{\alpha}_i$ por el polinomio $\frac{\tilde{l}}{\tilde{p}_i}$:

$$\tilde{\alpha}_i = \tilde{q}_i \frac{\tilde{l}}{\tilde{p}_i} + \tilde{r}_i.$$

Los restos \tilde{r}_i cumplen que $\deg(\tilde{r}_i) < \deg(\tilde{l}) - \deg(\tilde{p}_i)$. De la ecuación 4.1.0.1, obtenemos

$$\begin{aligned}\tilde{h}\tilde{g} &= \sum_{i=1}^n (\tilde{q}_i \frac{\tilde{l}}{\tilde{p}_i} \tilde{p}_i + \tilde{r}_i \tilde{p}_i) \\ &= \tilde{l} \left(\sum_{i=1}^n \tilde{q}_i \right) + \sum_{i=1}^n \tilde{r}_i \tilde{p}_i.\end{aligned}$$

Como $\deg(\tilde{h}\tilde{g}) < \deg(\tilde{l})$ y $\deg(\sum_{i=1}^n \tilde{r}_i \tilde{p}_i) < \deg(\tilde{l})$, resulta que debe ser $\sum_{i=1}^n \tilde{q}_i = 0$. De esta forma, $\tilde{h}\tilde{g} = \sum_{i=1}^n \tilde{r}_i \tilde{p}_i$ es una combinación que cumple lo buscado.

Podemos entonces recuperar la combinación (4.1.0.1) con los siguientes polinomios homogéneos

$$\begin{aligned}\alpha_i(x, y) &= y^{\deg(\tilde{l})-1-\deg(\tilde{p}_i)} \tilde{\alpha}_i\left(\frac{x}{y}\right), \\ p_i(x, y) &= y^{\deg(\tilde{p}_i)} \tilde{p}_i\left(\frac{x}{y}\right), \\ h(x, y) &= y^{\deg(\tilde{l})-1-\deg(\tilde{g})} \tilde{h}\left(\frac{x}{y}\right), \\ g(x, y) &= y^{\deg(\tilde{g})} \tilde{g}\left(\frac{x}{y}\right),\end{aligned}$$

de tal forma que $hg = \alpha_1 p_1 + \cdots + \alpha_n p_n$ y que el grado sea $\deg(l) - 1 = \deg(\tilde{l}) - 1$. Por último, observamos que g, p_1, \dots, p_n son efectivamente los polinomios originales del enunciado.

Probamos ahora el paso inductivo. Supongamos que $J_d \subset I_d$ y que $hg \in J_{d+1}$. Sean $h_0, \dots, h_{d+1} \in \mathbb{R}$ son los coeficientes del polinomio h que se escribe como

$$h = \sum_{i=0}^{d+1} h_i x^{d+1-i} y^i = h_0 x^{d+1} + \sum_{i=1}^d h_i x y x^{d-i} y^{i-1} + h_{d+1} y^{d+1}.$$

Entonces el producto hg puede escribirse como

$$hg = x(h_0 x^d g) + \sum_{i=1}^d xy(h_i x^{d-i} y^{i-1} g) + y(h_{d+1} y^d g). \quad (4.1.0.2)$$

Como los polinomios $h_0 x^d g, h_i x^{d-i} y^{i-1} g$ y $h_{d+1} y^d g \in J_d$ y suponemos que $J_d \subset I_d$, podemos reemplazarlos en (4.1.0.2) por combinaciones polinómicas de p_1, \dots, p_n :

$$hg = x \sum_{j=1}^n \alpha_j^0 p_j + \sum_{i=1}^d \sum_{j=1}^n xy \alpha_j^i p_j + y \sum_{j=1}^{d+1} \alpha_j^{d+1} p_j.$$

Como $\deg(\alpha_j^0) = d - \deg(p_j)$, entonces $\deg(x\alpha_j^0) = d + 1 - \deg(p_j)$. De manera análoga, cada factor $\alpha_j^i, \alpha_j^{d+1}$ tiene el grado buscado y vale que $hg \in I_{d+1}$. \square

Proposición 4.1.0.1. Sean (d, k, s) la arquitectura de una red de profundidad L de salto 1 y el mapa de parametrización μ . Sea k el tamaño del filtro de la convolución resultante de componer las capas de la red, que se calcula a partir de las coordenadas de k como en (1.1.0.10). Dado el vector de filtros $\theta = (w^1, \dots, w^L) \in \mathbb{R}^{k_1} \times \cdots \times \mathbb{R}^{k_L}$, consideramos $\bar{w} = \varphi(\mu(\theta)) \in \mathbb{R}^k$.

Como $\pi(\bar{w}) = \pi(w^1) \cdots \pi(w^L)$, podemos considerar el polinomio

$$g = \text{mcd}\left(\frac{\pi(\bar{w})}{\pi(w_1)}, \dots, \frac{\pi(\bar{w})}{\pi(w_L)}\right) \in \mathbb{R}[x, y].$$

Entonces, la imagen de la diferencial $d_\theta \mu$ en el punto θ se identifica con el conjunto de polinomios

$$J_{k-1}(\theta) := \{hg : h \in \mathbb{R}[x, y]_{k-1-\deg(g)}\} \subset \mathbb{R}[x, y]_{k-1}, \quad (4.1.0.3)$$

En particular, $d_\theta \mu$ es sobreyectiva si y solamente si $\text{mcd}(\pi(w^i), \pi(w^j)) = 1$ para cada par $i, j \in \{1, \dots, L\}$ distintos.

Demostración. Sea $\theta = (w^1, \dots, w^L)$. Para simplificar la notación, identificamos w^i con $\pi(w^i)$ para cada $i = 1, \dots, L$ y \bar{w} con $\pi(\bar{w})$.

4. Optimización

Daremos una caracterización de la imagen de $d_\theta \mu$ a través de la identificación con polinomios. Sea $M = \mathbb{R}_{k_1-1}[x, y] \times \cdots \times \mathbb{R}_{k_L-1}[x, y]$.

De esta manera,

$$d_\theta \mu : M \rightarrow \mathbb{R}_{k-1}[x, y].$$

Denotamos \dot{w}^i a las coordenadas del vector tangente correspondiente a la variable w^i .

Para comprender su imagen, consideramos una curva $\alpha : (-\varepsilon, \varepsilon) \rightarrow M$ en el espacio M definida como:

$$\begin{aligned} \alpha(t) &= (p_t^1(x, y), \dots, p_t^L(x, y)), \\ \alpha(0) &= (w^1, \dots, w^L) \text{ y } \dot{\alpha}(0) = (\dot{w}^1, \dots, \dot{w}^L). \end{aligned}$$

Derivamos, por regla del producto, como

$$\begin{aligned} (\mu \circ \alpha)'(0) &= (p_t^1(x, y) \cdots p_t^L(x, y))'|_{t=0} \\ &= \left[\sum_{i=1}^L \dot{p}_t^i \prod_{\substack{j=1 \\ j \neq i}}^L p_t^j(x, y) \right]_{t=0} \\ &= \sum_{i=1}^L \dot{w}^i \frac{\bar{w}}{w^i}. \end{aligned}$$

Luego, podemos concluir que

$$d_\theta \mu : (w^1, \dots, w^L) \mapsto \dot{w}^1 \frac{\bar{w}}{w^1} + \cdots + \dot{w}^L \frac{\bar{w}}{w^L}$$

y que la imagen de la diferencial se describe como el siguiente conjunto

$$\text{Im}(d_\theta \mu) = \left\{ \alpha_1 \frac{\bar{w}}{w^1} + \cdots + \alpha_L \frac{\bar{w}}{w^L} : \alpha_i \in \mathbb{R}[x, y], \deg(\alpha_i) = \deg(w^i) - 1 \right\}. \quad (4.1.0.4)$$

Aplicamos el Lema 4.1.0.1 con $d = k - 1$ y $g = \gcd(\frac{\bar{w}}{w^1}, \dots, \frac{\bar{w}}{w^L})$ y obtenemos

$$\text{Im}(d_\theta \mu) = \{hg : h \in \mathbb{R}[x, y], \deg(h) = k - 1 - \deg(g)\}.$$

A partir de esta caracterización, deducimos que $d_\theta \mu$ es sobreyectiva si y solamente si $g = 1$. Veamos que esta condición equivale a que $\gcd(w^i, w^j) = 1$ para $i \neq j$.

En efecto, cualquier factor común de w^i y w^j con $i \neq j$ es un factor de cada $\frac{\bar{w}}{w^m}$ con $m = 1, \dots, L$. Por lo tanto, como $g = \gcd(\frac{\bar{w}}{w^1}, \dots, \frac{\bar{w}}{w^L})$, si algún par de índices $i \neq j$ cumple que $\gcd(w^i, w^j) \neq 1$, entonces $g \neq 1$.

Ahora, si q es un factor común irreducible de $\frac{\bar{w}}{w^1}, \dots, \frac{\bar{w}}{w^L}$, entonces $q|w^i$ para algún $i = 1, \dots, L$. A su vez, $q|\frac{\bar{w}}{w^i}$ por lo que deducimos que $q|w^j$ para algún $j \neq i$. Entonces, si $\gcd(\frac{\bar{w}}{w^1}, \dots, \frac{\bar{w}}{w^L}) \neq 1$, se tiene que existen $i \neq j$ tales que $\gcd(w^i, w^j) \neq 1$. \square

Corolario 4.1.0.1. Sean $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ la arquitectura de una red de profundidad L de salto 1 y el mapa de parametrización μ . Sea k el tamaño del filtro de la convolución resultante de componer las capas de la red, que se calcula a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Dado el vector de filtros $\theta = (w^1, \dots, w^L) \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$, consideramos $\bar{w} = \varphi(\mu(\theta)) \in \mathbb{R}^k$. Vale que θ es un punto crítico espurio si y solamente si existen distintos $i, j \in \{1, \dots, L\}$ tales que $\pi(w^i)$ y $\pi(w^j)$ tienen alguna raíz en común.

Demostración. Como concluimos en la Observación (4.1.0.2), un punto crítico θ es espurio si y solamente si el rango de la diferencial $d_\theta \mu$ no es máximo. En la Proposición (4.1.0.1), demostramos que esto sucede si y solamente si existen distintos $i, j \in \{1, \dots, L\}$ tales que $\text{mcd}(\pi(w^i), \pi(w^j)) \neq 1$, es decir, tales que los polinomios $\pi(w^i)$ y $\pi(w^j)$ tienen alguna raíz en común. \square

En el caso de tener una función de pérdida convexa, este corolario nos permitirá concluir si un punto crítico $\theta = (w^1, \dots, w^L)$ es un mínimo global de la función objetivo \mathcal{L} a partir de las raíces de los polinomios $\pi(w^1), \dots, \pi(w^L)$.

Proposición 4.1.0.2. Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ la arquitectura de una red de profundidad L de salto 1. Sea k el tamaño del filtro de la convolución resultante de componer las capas de la red, que se calcula a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Consideramos una función ℓ suave y convexa en $\mathcal{M}_{((d_0, d_L), k, 1)}$.

Sea $\mathcal{L} = \ell \circ \mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ una función objetivo. Si $\theta = (w^1, \dots, w^L) \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ un punto crítico de \mathcal{L} , entonces vale alguna de las siguientes afirmaciones:

1. θ es un mínimo global de \mathcal{L} , o
2. para algún $i \neq j$ con $i, j \in \{1, \dots, L\}$, los polinomios homogéneos $\pi(w^i)$ y $\pi(w^j)$ alguna raíz en común.

Demostración. Si θ es un punto crítico puro, la hipótesis de convexidad sobre ℓ nos permite concluir que es un mínimo global.

Si θ no es un punto crítico puro, a partir de la Observación (4.1.0.2), sabemos que resulta ser un punto crítico espurio. A partir del Corolario (4.1.0.1), concluimos que deben existir distintos $i, j \in \{1, \dots, L\}$ tales que $\pi(w^i)$ y $\pi(w^j)$ tienen alguna raíz en común. \square

De la Proposición (4.1.0.1), podemos deducir que si tenemos un punto crítico θ de la función objetivo en el espacio de parámetros, todos los reescalamientos de θ también serán puntos críticos.

Lema 4.1.0.2. Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ la arquitectura de una red de profundidad L de salto 1 y μ el mapa de parametrización. Sea k el tamaño del filtro de las convoluciones del espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, que calculamos a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Sea $\mathcal{L} = \ell \circ \mu$ una función objetivo.

4. Optimización

Dada $\overline{W} \in \mathcal{M}_{d,k,s}$ de filtro $\overline{w} \in \mathbb{R}^k$, la fibra $\mu_{d,k,s}^{-1}(\overline{W})$ consiste en una cantidad finita de clases de equivalencia, dadas por la relación de la Definición 4.1.0.1.

Más aún, vale que si $\theta \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ es un punto crítico de \mathcal{L} , entonces todos los parámetros de su clase de equivalencia también lo son.

Demostración. Podemos comprender la fibra $\mu^{-1}(\overline{W}) = \{\theta = (w^1, \dots, w^L) : w^i \in \mathbb{R}^{k_i-1}, i = 1, \dots, L\}$ como las posibles maneras de describir al polinomio $\pi(\overline{w})$ como un producto de factores $\pi(w^i)$ de grado $k_i - 1$. Como $\pi(\overline{w})$ tiene una única factorización en factores irreducibles, escribirlo como tal producto es equivalente a combinar estos factores irreducibles para formar los polinomios $\pi(w^i)$ de grado $k_i - 1$ con $i = 1, \dots, L$. Hay finitas combinaciones. Más aún, si dos elementos $\theta = (w^1, \dots, w^L)$ y $\theta' = (v^1, \dots, v^L)$ son equivalentes, dan lugar al mismo polinomio:

$$\exists \kappa_1, \dots, \kappa_L \text{ tales que } w_i = \kappa_i v^i \text{ y } \prod_{i=1}^L \kappa_i = 1,$$

como $\pi(w^i) = \kappa_i \pi(v^i)$,

$$\prod_{i=1}^L \pi(w^i) = \prod_{i=1}^L \kappa_i \pi(v^i) = \prod_{i=1}^L \kappa_i \prod_{i=1}^L \pi(v^i) = \prod_{i=1}^L \pi(v^i).$$

Veamos ahora que si θ y θ' son equivalentes, entonces la imagen de la diferencial del mapa de parametrización μ coincide en ambos puntos. De esta manera,

$$d_\theta \mathcal{L} = d_{\overline{W}} \ell \circ d_\theta \mu = 0 \iff d_{\theta'} \mathcal{L} = d_{\overline{W}} \ell \circ d_{\theta'} \mu = 0.$$

En efecto, a partir de la Proposición (4.1.0.1), basta ver que

$$\gcd\left(\frac{\pi(\overline{w})}{\pi(w^1)}, \dots, \frac{\pi(\overline{w})}{\pi(w^L)}\right) = \gcd\left(\frac{\pi(\overline{w})}{\pi(v^1)}, \dots, \frac{\pi(\overline{w})}{\pi(v^L)}\right)$$

Esto se deduce del hecho que los polinomios $\pi(w^1), \dots, \pi(w^L)$ y $\pi(v^1), \dots, \pi(v^L)$ están conformados por los mismos factores irreducibles. \square

Ejemplo 4.1.0.3. Describimos las fibras correspondientes a distintos elementos del espacio de funciones de la red del Ejemplo (3.1.0.2) con $\mathbf{k} = (2, 2)$.

Si identificamos las matrices convolucionales con sus filtros, tenemos el mapa de parametrización

$$\begin{aligned} \mu : \mathbb{R}^2 \times \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ ((a, b), (c, d)) &\mapsto (ac, ad + bc, bd). \end{aligned}$$

En un punto $\theta = ((a, b), (c, d))$, podemos calcular su diferencial como

$$d_\theta \mu = \begin{bmatrix} c & 0 & a & 0 \\ d & c & b & a \\ 0 & d & 0 & b \end{bmatrix}.$$

Como vimos en el Capítulo 3, los elementos del interior del espacio de funciones se identifican con polinomios cuadráticos con dos raíces reales distintas

$$p(x, y) = (ax + by)(cx + dy).$$

Llamamos $w^1 = (a, b)$ al filtro que se identifica con el polinomio $p_1(x, y) = ax + by$ y $w^2 = (c, d)$ al filtro que se identifica con el polinomio $p_2(x, y) = cx + dy$. Para que el polinomio p tenga dos raíces reales distintas, los filtros no pueden ser múltiplos. Luego, la fibra $\mu^{-1}(\pi^{-1}(p))$ tiene cuatro componentes:

$$\begin{aligned} & \{(\kappa w^1, \frac{1}{\kappa} w^2) : \kappa > 0\}, \{(\kappa w^1, \frac{1}{\kappa} w^2) : \kappa < 0\}, \\ & \{(\kappa w^2, \frac{1}{\kappa} w^1) : \kappa > 0\}, \{(\kappa w^2, \frac{1}{\kappa} w^1) : \kappa < 0\}. \end{aligned}$$

En estos elementos, $d_\theta \mu$ tiene rango máximo.

Los elementos del borde del espacio de funciones se identifican con polinomios cuadráticos con una raíz real doble $p(x, y) = \pm(ex + fy)^2$ con e, f no ambos nulos. Llamamos $w' = (e, f)$. Luego, la fibra $\mu^{-1}(\pi^{-1}(p))$ tiene dos componentes $\{(\pm \kappa w', \frac{1}{\kappa} w') : \kappa > 0\}$ y $\{(\pm \kappa w', \frac{1}{\kappa} w') : \kappa < 0\}$. En este caso, los filtros son múltiplos y $d_\theta \mu$ no tiene rango máximo. Esto ilustra lo demostrado en la Proposición (4.1.0.1).

Dada una partición $\lambda = (\lambda_1, \dots, \lambda_r)$ de $k - 1$, recordamos la Definición (2.0.0.3) del conjunto $\Delta_\lambda \subset \mathbb{R}[x, y]_{k-1}$. Un polinomio p es un elemento del conjunto Δ_λ si puede escribirse como $p = q_1^{\lambda_1} \dots q_r^{\lambda_r}$ con $q_i \in \mathbb{C}[x, y]_1$.

Fijada una partición λ de $k - 1$, nos interesa además analizar la existencia de puntos críticos de la función de pérdida ℓ restringida a $\pi^{-1}(\Delta_\lambda)$. Es decir, de la función de pérdida restringida a convoluciones cuyo filtro se identifique con un polinomio de Δ_λ . Debemos considerar entonces una estructura diferencial en Δ_λ que desarrollaremos a partir de [17].

Definición 4.1.0.3. Definimos el conjunto $\mathbb{C}_{\text{sym}}^m$ como el conjunto de m -uplas de números complejos en el que identificamos dos elementos si solo difieren en una permutación de sus coordenadas. Denotamos $\langle \xi_1, \dots, \xi_m \rangle \in \mathbb{C}_{\text{sym}}^m$ con $\xi_1, \dots, \xi_m \in \mathbb{C}$.

Podemos pensar a un elemento $\xi = \langle \xi_1, \dots, \xi_m \rangle \in \mathbb{C}_{\text{sym}}^m$ como las raíces, listadas con multiplicidad, de un polinomio mónico $x^m + a_1 x^{m-1} + \dots + a_m \in \mathbb{C}[x]$. Como los coeficientes de este polinomio pueden calcularse a partir de sus raíces, definimos las *funciones simétricas* como

$$\sigma_1(\xi) = a_1, \dots, \sigma_m(\xi) = a_m.$$

Definimos la aplicación

$$\begin{aligned} \sigma : \mathbb{C}_{\text{sym}}^m &\rightarrow \mathbb{C}^m \\ \xi &\mapsto (\sigma_1(\xi), \dots, \sigma_m(\xi)), \end{aligned}$$

4. Optimización

en la que $(\sigma_1(\xi), \dots, \sigma_m(\xi))$ resulta ser la tira de coeficientes del polinomio mónico $(x - \xi_1) \dots (x - \xi_m)$. Esta aplicación es inyectiva y como \mathbb{C} es algebraicamente cerrado, resulta biyectiva. Definimos la inversa de σ como

$$\begin{aligned} \tau : \mathbb{C}^m &\longrightarrow \mathbb{C}_{\text{sym}}^m \\ (a_1, \dots, a_m) &\mapsto \xi \end{aligned}$$

con $\xi = \langle \xi_1, \dots, \xi_m \rangle$ las raíces del polinomio $x^m + a_1 x^{m-1} + \dots + a_m$.

En [17], se demuestra que σ y τ son homeomorfismos. Resultan difeomorfismos entre los conjuntos de puntos regulares de \mathbb{C}^m y de $\mathbb{C}_{\text{sym}}^m$.

Ejemplo 4.1.0.4. Podemos mirar a $\tau(\mathbb{R}^{k-1})$ dentro de $\mathbb{C}_{\text{sym}}^{k-1}$.

Por ejemplo, el conjunto $\mathbb{R}[x]_2$ de polinomios con coeficientes reales de grado 2 se identifica con el subconjunto propio de $\mathbb{C}_{\text{sym}}^2$

$$\tau(\mathbb{R}^2) = \{ \langle a_1 + ib_1, a_2 + ib_2 \rangle : b_1 = b_2 = 0 \text{ o } b_1 = -b_2, a_1 = a_2 \}.$$

Observación 4.1.0.4. En lugar de tomar polinomios homogéneos en dos variables, consideramos polinomios mónicos con coeficientes reales de grado $k - 1$ en una variable. Si miramos las tiras de coeficientes, podemos identificar

$$\mathbb{R}[x]_{k-1} \simeq \mathbb{R}^{k-1} \subset \mathbb{C}^{k-1} \xrightarrow{\tau} \mathbb{C}_{\text{sym}}^{k-1}.$$

Sea una partición λ de $k - 1$. Como $\Delta_\lambda \subset \mathbb{R}[x]_{k-1}$, $\tau(\Delta_\lambda) \subset \mathbb{C}_{\text{sym}}^{k-1}$. Diremos que un elemento $p \in \Delta_\lambda$ es un punto regular si $\tau(p)$ es un punto regular de $\tau(\Delta_\lambda)$. Si $\lambda = (\lambda_1, \dots, \lambda_r)$, los puntos regulares de Δ_λ son los polinomios p de este conjunto que pueden escribirse como el producto $p = q_1^{\lambda_1} \dots q_r^{\lambda_r}$ con $q_i \in \mathbb{C}[x]_1$ linealmente independientes dos a dos.

Ejemplo 4.1.0.5. Sea $\Delta_{(2,1)}$ el conjunto de polinomios de grado 3 que tienen una raíz doble y una raíz simple, no necesariamente distintas. Los puntos regulares de $\Delta_{(2,1)}$ son los polinomios que tienen exactamente una raíz doble y una simple distintas. Es decir, los polinomios que pueden escribirse como el producto $p = (x - \xi_1)^2(x - \xi_2)$ con $\xi_1, \xi_2 \in \mathbb{C}$ distintos.

Los Teoremas (4.1.0.1) y (4.1.0.2) de esta sección relacionan un punto crítico θ de \mathcal{L} tal que $\bar{\pi}(\mu(\theta))$ es un punto regular con los puntos críticos de ℓ restringida a $\bar{\pi}^{-1}(\Delta_\lambda)$. Estos resultados se desprenderán de estudiar la relación entre la imagen de la diferencial del mapa de parametrización μ en el punto θ y el espacio $T_{\bar{\pi}(\mu(\theta))}\Delta_\lambda$ tangente a Δ_λ en el punto regular $\bar{\pi}(\mu(\theta))$.

En efecto, como $\text{Im}(d_\theta \mu)$ y $T_{\bar{\pi}(\mu(\theta))}\Delta_\lambda$ se identifican con subconjuntos de $\mathbb{R}[x, y]_{k-1}$, cabe preguntarse qué relación existe entre ellos. En el caso en que $\bar{\pi}(\mu(\theta))$ sea un punto regular, podemos dar una caracterización de $T_{\bar{\pi}(\mu(\theta))}\Delta_\lambda$ en términos de su factorización.

Lema 4.1.0.3. Sean $k \in \mathbb{N}, k > 1$ y un filtro $\bar{w} \in \mathbb{R}^k$. Consideramos la factorización de $\pi(\bar{w}) = \prod_{i=1}^r q_i^{\lambda_i}$ en formas lineales complejas no equivalentes $q_i \in \mathbb{C}[x, y]_1$. Escribimos $\lambda = (\lambda_1, \dots, \lambda_r)$. El polinomio $\pi(\bar{w})$ es un punto regular de Δ_λ .

El espacio tangente a Δ_λ en $\pi(\bar{w})$ se describe como

$$T_{\pi(\bar{w})}\Delta_\lambda = \{h \prod_{i=1}^r q_i^{\lambda_i-1} : h \in \mathbb{R}[x, y]_r\} \subset \mathbb{R}[x, y]_{k-1}.$$

Demostración. Identificamos al filtro \bar{w} con el polinomio $\pi(\bar{w})$. Si $\bar{w} = \prod_{i=1}^r q_i^{\lambda_i}$ tiene coeficientes reales, entonces $\prod_{i=1}^r q_i^{\lambda_i-1}$ también. Esto vale pues si q_i tiene coeficientes complejos, entonces existe otro factor q_j que tiene coeficientes conjugados a los de q_i y están elevados al mismo exponente $\lambda_i = \lambda_j$. Entonces en el producto $\prod_{i=1}^r q_i^{\lambda_i-1}$, tienen también el mismo exponente $\lambda_i - 1 = \lambda_j - 1$.

De la factorización de $\bar{w} = q_1^{\lambda_1} \dots q_r^{\lambda_r}$, distinguimos los factores q_i según si tienen coeficientes reales o complejos. Sin pérdida de generalidad, $q_{i_1}, \dots, q_{i_n} \in \mathbb{R}[x, y]_1$. Agrupamos los factores restantes en pares conjugados para obtener formas reales cuadráticas irreducibles $t_{i_1}, \dots, t_{i_m} \in \mathbb{R}[x, y]_2$. Es decir, si $q_{i_j}, q_{i_{j+1}} \in \mathbb{C}[x, y]_1$ son conjugados, $\lambda_{i_j} = \lambda_{i_{j+1}}$, se obtiene la forma real cuadrática irreducible $t_{i_j} = q_{i_j} q_{i_{j+1}}$. Los enteros no negativos n, m cumplen que $r = n + 2m$. Podemos entonces escribir

$$\bar{w} = \prod_{j=1}^n q_{i_j}^{\lambda_{i_j}} \prod_{l=1}^m t_{i_l}^{\lambda_{i_l}}.$$

En un entorno de \bar{w} , Δ_λ es la imagen del morfismo

$$\psi : \mathbb{R}[x, y]_1 \times \dots \times \mathbb{R}[x, y]_1 \times \mathbb{R}[x, y]_2 \times \dots \times \mathbb{R}[x, y]_2 \rightarrow \mathbb{R}[x, y]_{k-1}$$

definido como

$$\psi : (\tilde{q}_{i_1}, \dots, \tilde{q}_{i_n}, \tilde{t}_{i_1}, \dots, \tilde{t}_{i_m}) \mapsto \prod_{j=1}^n \tilde{q}_{i_j}^{\lambda_{i_j}} \prod_{l=1}^m \tilde{t}_{i_l}^{\lambda_{i_l}}.$$

Entonces $T_{\bar{w}}\Delta_\lambda = d_{\bar{w}}\psi$.

Consideramos ahora $\tilde{g} = \gcd(\frac{\bar{w}}{q_{i_1}}, \dots, \frac{\bar{w}}{q_{i_n}}, \frac{\bar{w}}{t_{i_1}}, \dots, \frac{\bar{w}}{t_{i_m}})$. A partir de un argumento sobre las raíces de los polinomios a los que les buscamos el máximo común divisor, concluimos que $\tilde{g} = \prod_{i=1}^r q_i^{\lambda_i-1}$. Con el mismo argumento que se utiliza para demostrar la caracterización de la imagen de la diferencial del mapa de parametrización en la Proposición (4.1.0.3), deducimos que

$$T_{\bar{w}}\Delta_\lambda = \text{Im}(d_{\bar{w}}\varphi) = \{h\tilde{g} : h \in \mathbb{R}[x, y]_r\} = \{h \prod_{i=1}^r q_i^{\lambda_i-1} : h \in \mathbb{R}[x, y]_r\}.$$

□

4. Optimización

Corolario 4.1.0.2. Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ la arquitectura de una red de profundidad L de salto 1 y μ el mapa de parametrización. Calculamos k a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Sea $\theta \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ un vector de filtros y $\bar{w} = \varphi(\mu(\theta)) \in \mathbb{R}^k$ el filtro de la convolución resultante de componer las capas de la red. Consideramos el polinomio $\pi(\bar{w}) \in \mathbb{R}[x, y]_{k-1}$

Sea $\lambda = (\lambda_1, \dots, \lambda_r)$ una partición de $k - 1$ tal que el polinomio $\pi(\bar{w})$ es un punto regular de Δ_λ . Entonces, vale que

$$T_{\pi(\bar{w})}\Delta_\lambda \subseteq \text{Im}(d_\theta\mu).$$

Esta contención se entiende identificando cada conjunto con los subconjuntos de $\mathbb{R}[x, y]_{k-1}$ dados en el Lema (4.1.0.3) y la Proposición (4.1.0.1).

Demostración. Escribimos $\theta = (w^1, \dots, w^L)$. Identificamos a los filtros w^i con los polinomios $\pi(w^i)$, para cada $i = 1, \dots, L$ y al filtro \bar{w} con el polinomio $\pi(\bar{w}) = \pi(w^1) \dots \pi(w^L)$.

Por el Teorema (4.1.0.1), la imagen de $d_\theta\mu$ se identifica con el conjunto

$$J_{k-1}(\theta) = \{hg : h \in \mathbb{R}[x, y]_{k-1-\deg(g)}, \text{ con } g = \gcd\left(\frac{\bar{w}}{w^1}, \dots, \frac{\bar{w}}{w^L}\right)\}.$$

A su vez, por el Lema (4.1.0.3), el espacio tangente a Δ_λ en \bar{w} se describe como el conjunto

$$\{h \prod_{i=1}^r q_i^{\lambda_i-1} : h \in \mathbb{R}[x, y]_r\}.$$

Queremos ver que $g \mid \prod_{i=1}^r q_i^{\lambda_i-1}$ y así concluir que $T_{\bar{w}}\Delta_\lambda \subset J_{k-1}(\theta)$. Por definición, el polinomio $g = q_1^{\beta_1} \dots q_L^{\beta_L}$ con $\beta_1, \dots, \beta_L \in \mathbb{N}$. Supongamos que $q_i \mid w_j$ para algún par de índices i, j . Como q_i aparece con máximo exponente λ_i en la factorización de \bar{w} , y $q_i \neq q_j \forall i \neq j$, entonces $q_i^{\lambda_i}$ no divide a $\frac{\bar{w}}{w^j}$. Concluimos entonces que $\beta_i \leq \lambda_i - 1 \forall i = 1, \dots, L$ y así $g \mid \prod_{i=1}^r q_i^{\lambda_i-1}$. \square

Lema 4.1.0.4. Sea $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ una red de profundidad L de salto 1 y el mapa de parametrización μ . Calculamos k a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Sea $\bar{w} \in \mathbb{R}^k$ y λ una partición de $k - 1$.

Si $\pi(\bar{w})$ es un punto regular de Δ_λ , para cada $\theta = (w^1, \dots, w^L)$ tal que $\varphi(\mu(\theta)) = \bar{w}$, vale que

$$\text{Im}(d_\theta\mu) = T_{\pi(\bar{w})}\Delta_\lambda \iff \text{todo } \pi(w^i) \text{ tiene raíces simples.} \quad (4.1.0.5)$$

Esta igualdad se entiende identificando cada conjunto con los subconjuntos de $\mathbb{R}[x, y]_{k-1}$ dados en el Lema (4.1.0.3) y la Proposición (4.1.0.1).

Demostración. Identificamos a los filtros w^i con los polinomios $\pi(w^i)$, para cada $i = 1, \dots, L$ y al filtro \bar{w} con el polinomio $\pi(\bar{w}) = \pi(w^1) \dots \pi(w^L)$.

Consideramos la factorización de $\bar{w} = \prod_{i=1}^r q_i^{\lambda_i}$ en formas lineales complejas distintas $q_i \in \mathbb{C}[x, y]_1$. Como vimos en el Lema (4.1.0.3),

$$T_{\bar{w}}\Delta_\lambda = \{h \prod_{i=1}^r q_i^{\lambda_i-1} : h \in \mathbb{R}[x, y]_r\}.$$

Además, del Teorema (4.1.0.1), si $g = \gcd(\frac{\bar{w}}{w^1}, \dots, \frac{\bar{w}}{w^L})$, la imagen de la diferencial $d_\theta \mu$ se identifica con el conjunto

$$J_{k-1}(\theta) := \{hg : h \in \mathbb{R}[x, y]_{k-1-\deg(g)}\} \subset \mathbb{R}[x, y]_{k-1}.$$

Como vimos en el Corolario (4.1.0.2), $T_{\bar{w}}\Delta_\lambda \subseteq J_{k-1}(\theta)$. Para demostrar lo pedido, alcanza con ver que

$$J_{k-1}(\theta) \subseteq T_{\pi(\bar{w})}\Delta_\lambda \iff \text{todo } \pi(w^i) \text{ tiene raíces simples.}$$

Observamos que w^i tiene todas sus raíces simples si y solamente si $\prod_{i=1}^r q_i^{\lambda_i-1} \mid \frac{\bar{w}}{w^i}$. Esto sucede para todo $i = 1, \dots, L$ si y solamente si $\prod_{i=1}^r q_i^{\lambda_i-1} \mid g$.

Es decir, si $J_{k-1}(\theta) \subseteq T_{\pi(\bar{w})}\Delta_\lambda$. \square

En el siguiente teorema, veremos que a todo punto crítico del espacio de parámetros le corresponde un punto crítico en el espacio de funciones.

Si θ es un punto crítico de la función objetivo \mathcal{L} , la matriz $\mu(\theta) = \bar{W}$ tiene filtro $\bar{w} \in \mathbb{R}^k$ y le corresponde el polinomio $\pi(\bar{w})$. Entonces, la matriz \bar{W} resulta ser un punto crítico de la función de pérdida ℓ restringida al subconjunto del espacio de funciones cuyos filtros se identifican con polinomios que tienen la misma estructura de raíces que $\pi(\bar{w})$.

Teorema 4.1.0.1. Sean $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ la arquitectura de una red de profundidad L y de salto 1 y el mapa de parametrización μ . Sea k el tamaño del filtro de las convoluciones del espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, que calculamos a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Sea $\mathcal{L} = \ell \circ \mu$ una función objetivo.

Sea $\theta = (w^1, \dots, w^L) \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ un punto crítico de \mathcal{L} en el espacio de parámetros y $\bar{W} = \mu(\theta) \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ con filtro $\bar{w} \in \mathbb{R}^k$. Consideramos el polinomio $\bar{\pi}(\bar{W}) = \pi(\bar{w}) \in \mathbb{R}[x, y]_{k-1}$. Sea $\lambda = (\lambda_1, \dots, \lambda_r)$ una partición de $k-1$ tal que el polinomio $\pi(\bar{w})$ es un punto regular de Δ_λ .

Entonces, \bar{W} es un punto crítico de $\ell|_{\bar{\pi}^{-1}(\Delta_\lambda)}$.

Demostración. Si $\lambda = (1, \dots, 1)$, entonces $\pi(\bar{w})$ tiene $k-1$ raíces distintas. Por el Teorema (4.1.0.1), $d_\theta \mu$ tiene rango máximo, por lo que θ es un punto crítico puro.

4. Optimización

Resulta entonces un punto crítico de ℓ en $\mathbb{R}[x, y]_{k-1} = \Delta_\lambda$.

Supongamos ahora que $\lambda = (\lambda_1, \dots, \lambda_r)$ es un partición de $k - 1$ tal que, sin pérdida de generalidad, $\lambda_1 > 1$.

Como θ es un punto crítico de \mathcal{L} ,

$$0 = d_\theta \mathcal{L} = d_{\mu(\theta)} \ell \circ d_\theta \mu$$

y deducimos entonces que $(d_{\mu(\theta)} \ell)|_{\text{Im}(d_\theta \mu)} = 0$.

Para poder comprender esta restricción en el espacio de polinomios $\mathbb{R}[x, y]_{k-1}$, consideramos la función

$$\bar{\ell} = \ell \circ \bar{\pi}^{-1} : \mathbb{R}[x, y]_{k-1} \rightarrow \mathbb{R}.$$

Identificamos a $\mu(\theta) = \bar{W}$ con el polinomio $\bar{\pi}(\bar{W})$ y a $\text{Im}(d_\theta \mu)$ con el subconjunto $J_{k-1}(\theta)$ como consecuencia de la Proposición (4.1.0.1).

Por el Corolario (4.1.0.2), $T_{\bar{\pi}(\bar{W})} \Delta_\lambda \subseteq J_{k-1}(\theta)$. Luego, en el espacio de polinomios

$$(d_{\bar{\pi}(\bar{W})} \bar{\ell})|_{T_{\bar{\pi}(\bar{W})}} = 0$$

y entonces $\bar{\pi}(\bar{W})$ es punto crítico de ℓ restringida a Δ_λ . Concluimos entonces que $\bar{W} = \mu(\theta)$ es punto crítico de ℓ restringida a $\bar{\pi}^{-1}(\Delta_\lambda)$. \square

El siguiente teorema nos permitirá comprender qué elementos de la preimagen por el mapa de parametrización μ de un punto crítico de la función de pérdida en el espacio de funciones resultan ser también puntos críticos de la función objetivo en el espacio de parámetros.

Teorema 4.1.0.2. Sean $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ la arquitectura de una red de profundidad L y de salto 1 y el mapa de parametrización μ . Sea k el tamaño del filtro de las convoluciones del espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, que calculamos a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Sea $\mathcal{L} = \ell \circ \mu_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ una función objetivo.

Sean λ una partición de $k - 1$ y $\bar{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ tal que el polinomio $\bar{\pi}(\bar{W})$ es un punto regular de Δ_λ y \bar{W} es un punto crítico de $\ell|_{\bar{\pi}^{-1}(\Delta_\lambda)}$.

Si existe $\theta = (w^1, \dots, w^L)$ tal que $\mu(\theta) = \bar{W}$ y todos los polinomios $\pi(w^1), \dots, \pi(w^L)$ tienen raíces simples, entonces θ es un punto crítico de la función objetivo \mathcal{L} .

Demostración. Consideramos $\bar{\ell} = \ell \circ \bar{\pi}^{-1} : \mathbb{R}[x, y]_{k-1} \rightarrow \mathbb{R}$. Como \bar{W} es un punto crítico de ℓ restringida a $\bar{\pi}^{-1}(\Delta_\lambda)$, entonces $\bar{\pi}(\bar{W})$ es un punto crítico de $\bar{\ell}$ restringida a Δ_λ . Luego,

$$d_{\bar{\pi}(\bar{W})} \bar{\ell}|_{T_{\bar{\pi}(\bar{W})} \Delta_\lambda} = 0.$$

Como para todo $i = 1, \dots, L$ el polinomio $\pi(w^i)$ tiene raíces simples, por el Lema (4.1.0.4), $H(\theta) = T_{\bar{\pi}(\bar{w})}\Delta_\lambda$. Luego, como $\text{Im}(d_\theta\mu)$ se identifica con $H(\theta)$, deducimos que

$$(d_{\mu(\theta)}\ell)|_{\text{Im}(d_\theta\mu)} = 0.$$

Por lo tanto, como

$$d_\theta\mathcal{L} = d_{\mu(\theta)}\ell \circ d_\theta\mu = (d_{\mu(\theta)}\ell)|_{\text{Im}(d_\theta\mu)} = 0,$$

concluimos que θ es un punto crítico de \mathcal{L} . \square

Observación 4.1.0.5. Del Teorema 4.1.0.2, se sigue que un punto crítico de $\ell|_{\bar{\pi}^{-1}(\Delta_\lambda)}$ proviene de un punto crítico de \mathcal{L} si y solamente si $\bar{\pi}(\bar{w})$ puede escribirse como un producto, cuya estructura está determinada por la arquitectura de la red, en el que ningún factor tiene raíces múltiples. Más precisamente, la arquitectura determina la cantidad de factores L y, con la partición $(k_1 - 1, \dots, k_L - 1)$ de $k - 1$, el grado de cada factor. El hecho de que un polinomio con *rrmp* $(\rho|\gamma)$ pueda escribirse como un producto de polinomios de grado $k_1 - 1, \dots, k_L - 1$ en el que ningún factor tenga raíces múltiples es una propiedad puramente combinatoria. En efecto, hallar la escritura del polinomio como producto es equivalente a repartir ρ_i bolas de tamaño 1 y color i ($1 \leq i \leq r$) y γ_j bolas de tamaño 2 y color j ($1 \leq j \leq c$) en L cajas de tamaños $k_1 - 1, \dots, k_L - 1$ de tal forma que no haya bolas del mismo color en cada canasta.

Definición 4.1.0.4. Sea $(\gamma|\rho)$ un *rrmp* y una red de profundidad L con arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de salto 1. Sea k el tamaño del filtro de las matrices convolucionales que se obtienen al componer las capas de la red, se calcula a partir de las coordenadas de \mathbf{k} como en (1.1.0.10). Diremos que $(\gamma|\rho)$ es *compatible* con la arquitectura de la red si es posible escribir a un polinomio p de grado k y *rrmp* $(\gamma|\rho)$ como un producto de L factores de grados $k_1 - 1, \dots, k_L - 1$ en el que todos los factores tienen raíces simples.

Ejemplo 4.1.0.6. Comparamos las redes cuyas arquitecturas tienen como tamaños de filtros $\mathbf{k} = (3, 2, 2)$ y $\mathbf{k}' = (4, 2)$. El espacio de funciones correspondiente a cada arquitectura se identifica con el mismo espacio de polinomios $\mathbb{R}[x, y]_4$.

Consideremos el polinomio $\pi(\bar{w})$ con *rrmp* $(13|0)$. Su factorización es $\pi(\bar{w}) = p_1 p_2^3$ con $p_1, p_2 \in \mathbb{R}[x, y]_1$ linealmente independientes. Podemos escribirlo como un producto compatible con la arquitectura $\mathbf{k} = (3, 2, 2)$ como $\pi(\bar{w}) = (p_1 p_2) p_2 p_2$. En este caso, ningún factor tiene raíces múltiples. Al contrario, cualquier escritura de $\pi(\bar{w})$ acorde a la arquitectura $\mathbf{k}' = (4, 2)$ tendrá el factor cúbico asociado al filtro de tamaño 4 con raíces múltiples.

Listamos para cada arquitectura los *rrmp* $(\rho|\gamma)$ compatibles y proveemos una posible escritura como producto del polinomio. Los factores $p_i \in \mathbb{R}[x, y]_1$ linealmente independientes dos a dos y $q_1 \in \mathbb{R}[x, y]_2$:

4. Optimización

$\rho \gamma$	111 0	112 0	22 0	13 0	4 0	11 1	2 1	0 2	0 11
(3, 2, 2)	$(p_1 p_2) p_3 p_4$	$(p_1 p_2)(p_3)(p_4)$	$(p_1 p_2)(p_1)(p_2)$	$(p_1 p_2)(p_2)(p_2)$	-	$(q_1)(p_1)(p_2)$	$(q_1)(p_1)(p_1)$	-	-
(4, 2)	$(p_1 p_2 p_3) p_4$	$(p_1 p_2 p_3) p_3$	-	-	-	$(q_1 p_1)(p_2)$	$(p_1 q_1) p_1$	-	-

4.2. Dinámicas de entrenamiento

Consideramos la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L , su espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ y su mapa de parametrización μ . Sea el espacio de parámetros $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ y una función objetivo $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$, definida a partir de una función de pérdida $\ell : \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}} \rightarrow \mathbb{R}$, como en (4.0.0.1)

$$\mathcal{L} = \ell \circ \mu.$$

Una estrategia utilizada para aproximar mínimos locales de la función objetivo es la del *descenso del gradiente*. Es un algoritmo de optimización iterativo que, a partir de un elemento inicial $\theta_0 \in \mathcal{W}$, se desplazará hacia otro elemento en “la dirección” opuesta al gradiente de la función objetivo una distancia $\alpha > 0$:

$$\begin{aligned} \theta_1 &= \theta_0 - \alpha \nabla \mathcal{L}(\theta_0) \\ &\vdots \\ \theta_{n+1} &= \theta_n - \alpha \nabla \mathcal{L}(\theta_n). \end{aligned}$$

Esto define una curva en el espacio de parámetros. El objetivo es alcanzar un elemento cercano a algún mínimo local. Podemos también considerar una curva en el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ como

$$\overline{W}_n = \mu(\theta_n).$$

Definición 4.2.0.1. Sea la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L , su espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ y su mapa de parametrización μ . Sea el espacio de parámetros $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ y una función objetivo $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$ como en la Definición 4.0.0.1. Sea $I = [0, N] \subset \mathbb{R}$ un intervalo.

Una *curva integral* en el espacio de parámetros es una curva tal que

$$\theta(t) : I \rightarrow \mathcal{W}; \quad \dot{\theta}(t) = -\nabla(\mathcal{L}(\theta(t))).$$

Le corresponde una curva $\overline{W} : I \rightarrow \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ en el espacio de funciones de la forma

$$\overline{W}(t) = \mu(\theta(t)).$$

Como vimos en la sección anterior, un mismo elemento $\overline{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ puede ser parametrizado por distintos elementos del espacio de parámetros. En la Definición (4.1.0.2), llamamos a este conjunto la fibra de \overline{W} o de su filtro \overline{w} . Una pregunta que surge en este marco es cómo se puede recuperar entonces una curva en el espacio de parámetros a partir de una curva definida en el espacio de funciones.

Observación 4.2.0.1. Sea el mapa de parametrización de la red $\mu : \mathcal{W} \rightarrow \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$. El espacio de funciones es un subconjunto de $\mathbb{R}^{d_L \times d_0}$, conformado por matrices convolucionales con tamaño de filtro k que se calcula como en la Proposición 1.1.0.1. Sea la función objetivo $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$ que se calcula a partir de la función de pérdida como

$$\mathcal{L} = \ell \circ \mu.$$

Recordamos que $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$. En esta sección, denotaremos $\nabla \mathcal{L} \in \mathbb{R}^{(k_1 + \dots + k_L) \times 1}$ al gradiente de la función objetivo, $\nabla \ell \in \mathbb{R}^{d_0 d_L \times 1}$ de pérdida y $\text{Jac}_\mu \in \mathbb{R}^{d_0 d_L \times (k_1 + \dots + k_L)}$ a la matriz jacobiana del mapa de parametrización. Por la regla de la cadena, en un elemento $\theta \in \mathcal{W}$ vale que

$$\nabla \mathcal{L}(\theta) = \text{Jac}_\mu(\theta)^T \nabla \ell(\mu(\theta)). \quad (4.2.0.1)$$

Consideramos una curva integral $\theta(t)$ en el espacio de parámetros y la curva $\overline{W}(t) = \mu(\theta(t))$ que le corresponde en el espacio de funciones como en la Definición (4.2.0.1). Si derivamos la curva en el espacio de funciones,

$$\begin{aligned} \frac{d}{dt} \overline{W}(t) &= \frac{d}{dt} \mu(\theta(t)) \\ &= \text{Jac}_\mu(\theta(t)) \cdot -\nabla \mathcal{L}(\theta(t)) \\ &= -\text{Jac}_\mu(\theta(t)) \cdot \text{Jac}_\mu(\theta(t))^T \nabla \ell(\overline{W}(t)). \end{aligned}$$

Llamamos $K(\theta(t)) := \text{Jac}_\mu(\theta(t)) \text{Jac}_\mu(\theta(t))^T$, conocida como *neural tangent kernel* en el análisis de las redes neuronales. Esta matriz depende del valor de los parámetros $\theta(t)$ y no del filtro correspondiente en el espacio de funciones $\mu(\theta(t))$. Por lo tanto, una curva en el espacio de funciones no depende únicamente del elemento \overline{W} en que se inicia sino también de cómo está parametrizado este elemento. Una de las dificultades del problema es entonces que los parámetros de la misma fibra pueden dar lugar a distintos *neural tangent kernels* y por lo tanto a distintas curvas que se inician en el mismo elemento del espacio de funciones.

Los resultados demostrados a continuación permitirán, en el caso de redes de salto 1, recuperar la curva correspondiente en el espacio de parámetros a partir de una curva definida en el espacio de funciones.

En primer lugar, demostraremos la existencia de un invariante para una curva en el espacio de parámetros.

Proposición 4.2.0.1. *Sea la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L de salto 1, su espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ y su mapa de parametrización μ . Sea el espacio de parámetros $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ y una función objetivo $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$ como en (4.0.0.1). Consideramos $\theta(t) = (w^1(t), \dots, w^L(t))$ una curva integral en el espacio de parámetros definida como en (4.2.0.1). Entonces las cantidades*

$$\delta_{ij}(t) = \|w^i(t)\|^2 - \|w^j(t)\|^2, \quad i, j \in \{1, \dots, L\}$$

se mantienen constantes para todo valor de $t \in I$.

4. Optimización

Demostración. Para cada $i, j \in \{1, \dots, L\}$ distintos, definimos la función

$$\begin{aligned}\nu_{i,j} : \mathcal{W} \times \mathbb{R} &\rightarrow \mathcal{W}, \\ \nu_{i,j}(\theta, \kappa) &= (w^1, \dots, \kappa w^i, \dots, \frac{1}{\kappa} w^j, \dots, w^L).\end{aligned}$$

Como el mapa de parametrización μ es multilineal en los filtros, para cualquier $\kappa \in \mathbb{R}$

$$\mu(\nu_{i,j}(\theta, \kappa)) = \mu(\theta) \implies \mathcal{L}(\nu_{i,j}(\theta, \kappa)) = \mathcal{L}(\theta).$$

resulta que $\mathcal{L}(\nu_{i,j}(\theta, \kappa))$ no depende de κ . Luego,

$$\frac{d}{d\kappa} \mathcal{L}(\nu_{i,j}(\theta, \kappa)) \equiv 0.$$

Para poder calcular esta derivada, obtenemos primero

$$\begin{aligned}\frac{d}{d\kappa} \nu_{i,j}(\theta, \kappa)(w^1, \dots, w^L) &= \\ (0, \dots, \kappa w^i, \dots, \frac{-1}{\kappa^2} w^j, \dots, 0) &\in \mathbb{R}^{1 \times (k_1 + \dots + k_L)}.\end{aligned}$$

En particular, calculamos para $\kappa = 1$, se tiene

$$0 = \left[\frac{d}{d\kappa} \mathcal{L}(\nu_{i,j}(\theta, \kappa)) \right]_{\kappa=1} \quad (4.2.0.2)$$

$$= (0, \dots, w^i, \dots, -w^j, \dots, 0) \cdot \nabla \mathcal{L}(\theta). \quad (4.2.0.3)$$

Queremos expresar a las cantidades δ_{ij} como funciones que dependen de los filtros (w^1, \dots, w^L) . Definimos para cada $i, j \in \{1, \dots, L\}$ distintos,

$$\begin{aligned}\eta_{ij} : \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L} &\rightarrow \mathbb{R} \\ (w^1, \dots, w^L) &\mapsto \|w^i\|^2 - \|w^j\|^2.\end{aligned}$$

De esta manera, $\delta_{ij}(t) = \eta_{ij} \circ \theta(t) : I \subset \mathbb{R} \rightarrow \mathbb{R}$. Para poder derivar estas funciones, calculamos

$$\nabla \eta_{ij} = 2 (0, \dots, w^i, \dots, -w^j, \dots, 0)^T \in \mathbb{R}^{(k_1 + \dots + k_L) \times 1}.$$

Para cada elección de distintos $i, j \in \{1, \dots, L\}$, derivamos

$$\frac{d}{dt} \delta_{ij}(t) = \frac{d}{d\theta} \delta_{ij} \cdot \dot{\theta}(t) = 2 (0, \dots, w^i, \dots, -w^j, \dots, 0) \cdot (-\nabla \mathcal{L}(\theta(t))) = 0$$

por lo calculado en (4.2.0.2). Concluimos que d_{ij} es constante para todo valor de t . \square

Sea $\overline{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ un elemento del espacio de funciones y un vector de filtros θ que lo parametriza. Como se discute en el Lema 4.1.0.2, la fibra $\mu^{-1}(\overline{W})$ es un conjunto de finitas clases de equivalencia. Cada clase consiste en reescalamientos de un vector

de filtros representante. Hay infinitos reescalamientos de este representante que se corresponden con elecciones de valores $\kappa_1, \dots, \kappa_L \in \mathbb{R}$ como en la Definición 4.1.0.1. No es posible saber a priori qué elemento de la clase de equivalencia parametriza a \overline{W} .

Dada una curva integral $\overline{W}(t)$ en el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$, nos gustaría comprender qué vectores de parámetros en \mathcal{W} parametrizan cada punto de la curva. Si conocemos el vector de filtros $\theta(0)$ que parametriza al elemento $\overline{W}(0)$ en que se inicia la curva, la Proposición 4.2.0.1 nos permite determinar los valores $\kappa_1, \dots, \kappa_L$ y así recuperar la curva integral $\theta(t)$ en el espacio de parámetros.

Corolario 4.2.0.1. *Sea la arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ de una red de profundidad L de salto 1 y el mapa de parametrización μ . Sea el espacio de parámetros $\mathcal{W} = \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ y una función objetivo $\mathcal{L} : \mathcal{W} \rightarrow \mathbb{R}$ como en (4.0.0.1).*

Sean $\theta = (w^1, \dots, w^L), \theta' = (v^1, \dots, v^L) \in \mathcal{W}$. Para cada par de índices distintos $i, j = 1, \dots, L$, calculamos las cantidades

$$\begin{aligned}\delta_{ij}(t) &= \|w^i(t)\|^2 - \|w^j(t)\|^2 \\ \delta'_{ij}(t) &= \|v^i(t)\|^2 - \|v^j(t)\|^2.\end{aligned}$$

El conjunto de vectores de filtros $\theta' \in \mathcal{W}$ tales que $\mu(\theta') = \mu(\theta)$ y $\delta_{i,i+1} = \delta'_{i,i+1}$ para todo $i = 1, \dots, L-1$ es finito. Es decir, hay finitos reescalamientos de θ tales que las cantidades $\delta_{i,i+1}$ y $\delta'_{i,i+1}$ coinciden para todo $i = 1, \dots, L-1$.

Demostración. Sean θ y θ' como en el enunciado y los polinomios $q_i = \pi(w^i), q'_i = \pi(v^i)$ para $i = 1, \dots, L$. Llamamos $p = \bar{\pi}(\mu(\theta)) = q_1 \dots q_L = q'_1 \dots q'_L$.

Los vectores de filtros que están en la misma fibra que θ se identifican el polinomios obtenidos al repartir las raíces de p en L factores de grados $k_1 - 1, \dots, k_L - 1$. Esta acción arroja finitos resultados posibles. Dentro de la fibra, hay infinitos parámetros θ' equivalentes a θ que se corresponden con una elección de valores $\kappa_1, \dots, \kappa_L \in \mathbb{R}$ tales que

$$\begin{aligned}\prod_{i=1}^L \kappa_i &= 1, \\ q'_i &= \kappa_i q_i, \quad \forall i = 1, \dots, L.\end{aligned}$$

Debemos ver que si $\delta_{i,i+1} = \delta'_{i,i+1} \quad \forall i = 1, \dots, L-1$, entonces hay finitas elecciones posibles para $\kappa_1, \dots, \kappa_L$.

En efecto, definimos $\beta_i = \|\kappa_i w^i\|^2 = \kappa_i^2 \|w^i\|^2$ y obtenemos las siguientes ecuaciones

$$\beta_{i+1} - \beta_i = \delta_{i+1,i} \quad i = 1, \dots, L-1 \quad (4.2.0.4)$$

$$\prod_{i=1}^L \beta_i = \prod_{i=1}^L \|w^i\|^2. \quad (4.2.0.5)$$

4. Optimización

En particular, todos los valores de β_i están determinados por β_1

$$\beta_{i+1} = \beta_1 + \delta_{2,1} + \cdots + \delta_{i+1,i}$$

y esto nos permite reescribir la ecuación (4.2.0.5) como

$$\beta_1(\beta_1 + \delta_{2,1}) \cdots (\beta_1 + \delta_{2,1} + \cdots + \delta_{L,L-1}) - \prod_{i=1}^L \|w^i\|^2 = 0 \quad (4.2.0.6)$$

Como hay finitas soluciones para esta ecuación, deducimos que hay finitas elecciones posibles para $\kappa_1, \dots, \kappa_L$. \square

Veamos cómo aplicarlo en el siguiente ejemplo.

Ejemplo 4.2.0.1. Consideremos una arquitectura con $\mathbf{k} = (3, 2)$. El espacio de funciones se identifica con polinomios de grado 3 que se escriben como el producto de un polinomio cuadrático y uno lineal

$$c_0x^3 + c_1x^2y + c_2xy^2 + c_3y^3 = (a_0x^2 + a_1xy + a_2y)(b_0x + b_1y).$$

Describimos al mapa de parametrización como

$$\mu : ((a_0, a_1, a_2), (b_0, b_1)) \mapsto (c_0, c_1, c_2, c_3) = (a_0b_0, a_0b_1 + a_1b_0, a_1b_1 + a_2b_0, a_2b_1).$$

Calculamos su matriz jacobiana

$$\text{Jac}_\mu((a_0, a_1, a_2), (b_0, b_1)) = \begin{bmatrix} b_0 & 0 & 0 & a_0 & 0 \\ b_1 & b_0 & 0 & a_1 & a_0 \\ 0 & b_1 & b_0 & a_2 & a_1 \\ 0 & 0 & b_1 & 0 & a_2 \end{bmatrix} \in \mathbb{R}^{4 \times 5}.$$

La matriz $K = \text{Jac}_\mu \text{Jac}_\mu^T$ puede escribirse como la suma de la matriz K_A que depende únicamente de los coeficientes del polinomio cuadrático y de la matriz K_B que depende únicamente de los coeficientes del polinomio lineal:

$$K_A = \begin{bmatrix} a_0^2 & a_1a_0 & a_0a_2 & 0 \\ a_1a_0 & a_1^2 + a_0^2 & a_2a_1 + a_1a_0 & a_0a_2 \\ a_0a_2 & a_2a_1 + a_1a_0 & a_2^2 + a_1^2 & a_2a_1 \\ 0 & a_0a_2 & a_2a_1 & a_2^2 \end{bmatrix}, K_B = \begin{bmatrix} b_0^2 & b_0b_1 & 0 & 0 \\ b_0b_1 & b_0^2 + b_1^2 & b_0b_1 & 0 \\ 0 & b_0b_1 & b_0^2 + b_1^2 & b_0b_1 \\ 0 & 0 & b_0b_1 & b_1^2 \end{bmatrix}.$$

La matrix K depende claramente de los valores del punto considerado en el espacio de parámetros. Podemos recuperar estos valores a partir del filtro (c_0, c_1, c_2, c_3) del espacio de funciones y de la cantidad $\delta = \|(a_0, a_1, a_2)\|^2 - \|(b_0, b_1)\|^2$ que queda fija desde el inicio del recorrido de la curva de descenso del gradiente.

4.3. Optimización de la función de error cuadrático

Para esto, consideramos $\gamma(a_0, a_1, a_2)$ y $\frac{1}{\gamma}(b_0, b_1)$ para cualquier valor de γ no nulo. Entonces

$$K((a_0, a_1, a_2), (b_0, b_1)) = \gamma^2 K_A + \frac{1}{\gamma} K_B \quad (4.2.0.7)$$

$$\|\gamma(a_0, a_1, a_2)\|^2 - \left\| \frac{1}{\gamma}(b_0, b_1) \right\|^2 = \delta. \quad (4.2.0.8)$$

Como en la demostración anterior, definimos $\beta = \left\| \frac{1}{\gamma}(b_0, b_1) \right\|^2$ y $H = \|(a_0, a_1, a_2)\|^2 \|(b_0, b_1)\|^2$. Si multiplicamos la ecuación (4.2.0.8) por β , obtenemos las siguientes restricciones para β

$$\beta^2 + \delta\beta - H = 0, \quad \beta > 0, \quad \beta + \delta > 0, \quad H > 0.$$

Notamos que es la misma ecuación que (4.2.0.6). Quedan las soluciones

$$\beta = \frac{-\delta + \sqrt{\delta^2 + 4H}}{2}, \quad \gamma^2 = \frac{2 \|(b_0, b_1)\|^2}{-\delta + \sqrt{\delta^2 + 4H}}.$$

Con esto, se pueden recuperar los valores $((a_0, a_1, a_2), (b_0, b_1))$.

4.3. Optimización de la función de error cuadrático

En esta sección, estudiaremos la existencia de puntos críticos y mínimos para la función de error cuadrático.

4.3.1. La función de error cuadrático

A lo largo de esta subsección, las matrices $X \in \mathbb{R}^{d_0 \times N}$ e $Y \in \mathbb{R}^{d_L \times N}$ representarán los N datos del conjunto de entrenamiento.

Para poder definir la función de error cuadrático, daremos unas definiciones previas. sobre productos internos y normas en el espacio de matrices.

Definición 4.3.1.1. Sean $n, m \in \mathbb{N}$ y las matrices $A, B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{m \times m}$. Se definen

$$\begin{aligned} \langle A, B \rangle &= \text{Tr}(AB^T) \\ \langle A, B \rangle_C &= \text{Tr}(ACB^T) \\ \|A\|_C^2 &= \text{Tr}(ACA^T). \end{aligned}$$

Definición 4.3.1.2. Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}$, $Y \in \mathbb{R}^{d_L \times N}$ los datos de entrenamiento y $W \in \mathbb{R}^{d_0 \times d_L}$. Asumimos que XX^T tiene rango máximo.

La *función de error cuadrático* se define como

$$\ell(W) = \ell_{X,Y}(W) = \|WX - Y\|^2.$$

4. Optimización

Lema 4.3.1.1. Consideramos una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$. Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}$, $Y \in \mathbb{R}^{d_L \times N}$ los datos de entrenamiento.

Sea $U \in \mathbb{R}^{d_L \times d_0}$ la solución al problema de optimización sin restricciones

$$U = \min_{V \in \mathbb{R}^{d_L \times d_0}} \|VX - Y\|^2 = YX^T(XX^T)^{-1}.$$

Entonces, vale que

$$\min_{\bar{W} \in \mathcal{M}} \ell(\bar{W}) = \min_{\bar{W} \in \mathcal{M}} \|U - \bar{W}\|_{XX^T}^2. \quad (4.3.1.1)$$

Demostración. Para demostrar (4.3.1.1), reescribimos la norma como un producto interno

$$\|WX - Y\|^2 = \langle WX - Y, WX - Y \rangle.$$

Lo desarrollamos linealmente

$$\langle Y, Y \rangle - 2\langle WX, Y \rangle + \langle WX, WX \rangle = \langle Y, Y \rangle - 2\langle W, YX^T \rangle + \langle WX, WX \rangle$$

Como $\langle Y, Y \rangle$ es constante y asumimos que XX^T tiene rango máximo,

$$\begin{aligned} \langle Y, Y \rangle - 2\langle WX, Y \rangle + \langle WX, WX \rangle &= \text{const} + 2\langle W(XX^T), YX^T(XX^T)^{-1} \rangle + \langle WX, WX \rangle \\ &= \text{const} + 2\langle W(XX^T), YX^T(XX^T)^{-1} \rangle + \langle WXX^T, W \rangle \end{aligned}$$

A partir de la definición de U , reescribimos la expresión obtenida como

$$\begin{aligned} &= \text{const} + 2\langle W(XX^T), U \rangle + \langle WXX^T, W \rangle \\ &= \text{const} + 2\langle W, U \rangle_{XX^T} + \langle W, W \rangle_{XX^T} \\ &= \text{const} + \|W - U\|_{XX^T}^2. \end{aligned}$$

□

Observación 4.3.1.1. Consideramos una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$.

Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}$, $Y \in \mathbb{R}^{d_L \times N}$ los datos de entrenamiento. Podemos hallar

$$U = \min_{V \in \mathbb{R}^{d_L \times d_0}} \|VX - Y\|^2 = YX^T(XX^T)^{-1}.$$

En el Lema (4.3.1.1), vimos que optimizar la función de error cuadrático sobre el espacio de funciones $\mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ es equivalente a buscar

$$\min_{\bar{W} \in \mathcal{M}} \|U - \bar{W}\|_{XX^T}^2.$$

4.3. Optimización de la función de error cuadrático

Si $U \in \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$, entonces el mínimo buscado es U . Si no, optimizamos la función de pérdida sobre el borde del espacio de funciones:

$$\min_{\bar{W} \in \partial \mathcal{M}} \ell(\bar{W}).$$

Como será conveniente trabajar con matrices convolucionales, consideramos \bar{U} la proyección ortogonal de U con el producto interno $\langle -, - \rangle_{XX^T}$ al espacio $\mathcal{M}_{(d_0, d_L), k, s}$. Vale entonces

$$\min_{\bar{W} \in \partial \mathcal{M}} \|U - \bar{W}\|_{XX^T}^2 = V \iff \min_{\bar{W} \in \partial \mathcal{M}} \|\bar{U} - \bar{W}\|_{XX^T}^2 = V.$$

Definición 4.3.1.3. Consideramos una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$.

Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}, Y \in \mathbb{R}^{d_L \times N}$ los datos de entrenamiento. Sea

$$U = \min_{V \in \mathbb{R}^{d_L \times d_0}} \|VX - Y\|^2 = YX^T(XX^T)^{-1}.$$

Definimos la matriz $\bar{U} \in \mathcal{M}_{(d_0, d_L), k, s}$ de filtro $\bar{u} \in \mathbb{R}^k$ como la proyección ortogonal de U con el producto interno $\langle -, - \rangle_{XX^T}$ al espacio $\mathcal{M}_{(d_0, d_L), k, s}$. Llamamos *filtro objetivo* a \bar{u} .

A partir de esta observación, podemos dar una escritura equivalente de la función de error cuadrático a partir de los filtros de las matrices convolucionales involucradas.

Lema 4.3.1.2. Sea una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$.

Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}, Y \in \mathbb{R}^{d_L \times N}$ los datos de entrenamiento. Sea \bar{U} la matriz convolucional de salto s y filtro $\bar{u} \in \mathbb{R}^k$ de la Definición 4.3.1.3. Definimos el operador $\mathcal{T}_s : \mathbb{R}^{d_0 \times d_0} \rightarrow \mathbb{R}^{k \times k}$ como

$$(\mathcal{T}_s)_{i,j} = \sum_{m=0}^{d_L-1} M_{i+sm, j+sm}.$$

Dada $\bar{W} \in \mathcal{M}_{\mathbf{d},\mathbf{k},\mathbf{s}}$ de filtro $\bar{w} \in \mathbb{R}^k$, la función de error cuadrático puede escribirse en función de los filtros como

$$\ell(\bar{w}) = (\bar{w} - \bar{u})^T \mathcal{T}_s(XX^T)(\bar{w} - \bar{u}).$$

Demostración. Dada una matriz M , notaremos $M_{m,:}$ a su m -ésima columna.

4. Optimización

Reescribimos la función de error cuadrático como

$$\begin{aligned}\ell(\bar{W}) &= \|\bar{W} - \bar{U}\|_{XX^T}^2 = \text{Tr}[(\bar{W} - \bar{U})XX^T(\bar{W} - \bar{U})^T] \\ &= \sum_{m=0}^{d_L-1} [(\bar{W} - \bar{U})_{m,:}XX^T(\bar{W} - \bar{U})_{m,:}^T].\end{aligned}$$

Como la m -ésima columna de $\bar{W} - \bar{U}$ se obtiene desplazando sm posiciones a la derecha el vector $\bar{w} - \bar{u}$, podemos reescribir

$$\ell(\bar{w}) = (\bar{w} - \bar{u})^T \mathcal{T}_s(XX^T)(\bar{w} - \bar{u}).$$

□

Si los datos de entrenamiento tienen una estructura particular, la expresión anterior permite describir a la función de error cuadrático de forma simplificada como la distancia entre filtros.

Lema 4.3.1.3. *Sea una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$. Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$.*

Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}$, $Y \in \mathbb{R}^{d_L \times N}$ los datos de entrenamiento. Sea \bar{U} la matriz convolucional de salto s y filtro $\bar{u} \in \mathbb{R}^k$ de la definición 4.3.1.3.

Si XX^T es un múltiplo de la matriz identidad, para cualquier salto s la función de error cuadrático definida sobre los filtros \bar{w} y u es equivalente a su distancia, es decir,

$$\ell(\bar{w}) \propto \|\bar{w} - u\|^2.$$

Demostración. Como XX^T es un múltiplo de la matriz identidad I , podemos denotar $XX^T = \lambda I$ con $\lambda \in \mathbb{R}$. Luego, $\tau(XX^T) = d_L XX^T = d_L \lambda I$. Por lo que concluimos, a partir del lema 4.3.1.2

$$\ell(\bar{w}) = (\bar{w} - u)d_L \lambda (\bar{w} - u)^T = d_L \lambda \|\bar{w} - u\|^2.$$

□

4.3.2. Puntos críticos

Como vimos en los resultados anteriores, el tipo de datos con el que contemos para entrenar al modelo va a influenciar el comportamiento de la función de error cuadrático. Bajo buenas hipótesis sobre estos datos, podremos garantizar que no haya puntos críticos espurios. Es decir, que todo punto crítico de una función objetivo genérica se corresponda con un punto crítico de la función genérica de pérdida correspondiente. Para eso, veremos que si se tiene un punto crítico θ de la función objetivo genérica, su identificación polinomial $\bar{\pi}(\mu(\theta))$ puede escribirse como un producto, cuya estructura está determinada por la red, en el que todos los factores tienen raíces simples.

Teorema 4.3.2.1. *Dada una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ y el mapa de parametrización μ . Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$. Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}$, $Y \in \mathbb{R}^{d_L \times N}$ datos de entrenamiento genéricos y $W \in \mathbb{R}^{d_0 \times d_L}$. Sea ℓ la función de error cuadrático de la Definición 4.3.1.2 y $\mathcal{L} = \ell \circ \mu$.*

Si XX^T tiene rango máximo, entonces todo punto crítico $\theta = (w^1, \dots, w^L) \in \mathbb{R}^{k_1} \times \dots \times \mathbb{R}^{k_L}$ de \mathcal{L} es puro.

Demostración. En el Lema 4.3.1.2, vimos que la función de error cuadrático puede escribirse como

$$\ell(\bar{w}) = (\bar{w} - \bar{u})^T \tau(XX^T)(\bar{w} - \bar{u})$$

con $\bar{u} \in \mathbb{R}^k$ el filtro objetivo definido en la Definición (4.3.1.3). En esta escritura, el producto interno definido depende únicamente de la matriz de datos de entrada X . La matriz de datos de salida Y solo influye en el cómputo del filtro objetivo \bar{u} . Luego, como podemos combinar matrices de datos de entrada y de salida, cualquier combinación de un producto interno arbitrario que denotamos $\langle \cdot, \cdot \rangle_\Sigma$ y de un filtro objetivo genérico \bar{u} da lugar a una función de pérdida $\ell_{\Sigma, \bar{u}}$. A su vez, esto da lugar a la función objetivo

$$\mathcal{L}_{\Sigma, \bar{u}} = \ell_{\Sigma, \bar{u}} \circ \mu \quad (4.3.2.1)$$

Vamos a ver que para cualquier producto interno $\langle \cdot, \cdot \rangle_\Sigma$ arbitrario y cualquier filtro objetivo \bar{u} genérico, cualquier punto crítico $\theta = (w^1, \dots, w^L)$ de $\mathcal{L}_{\Sigma, \bar{u}}$ satisface que los polinomios $\pi(w^1), \dots, \pi(w^L)$ tienen raíces simples. Por el Lema 4.1.0.4, esto quiere decir que el punto crítico es puro.

Fijamos un producto interno $\langle \cdot, \cdot \rangle_\Sigma$, para cada filtro objetivo u , se define $\mathcal{L}_{\bar{u}}$ como en (4.3.2.1). Para cada $i = 1, \dots, L$, definimos

$$Z_i = \{(\theta, \bar{w}, \bar{u}) : \theta \in \text{Crit}(\mathcal{L}_{\bar{u}}), \mu(\theta) = \bar{w}, w^i \text{ tiene raíces repetidas}\}.$$

Se trata de un conjunto algebraico. Vamos a demostrar que la proyección a la tercera coordenada $\text{pr}_3(Z_i)$ tiene dimensión menor estricta que k para todo $i = 1, \dots, L$. Esto quiere decir que el conjunto de filtros objetivo \bar{u} para los cuales cualquier punto crítico correspondiente tiene algún filtro w^i tal que el polinomio $\pi(w^i)$ tiene raíces múltiples es un conjunto de codimensión positiva.

Más aún, consideramos para cada λ una partición de $k - 1$

$$Z_{i, \lambda} = \{(\theta, \bar{w}, \bar{u}) : (\theta, \bar{w}, \bar{u}) \in Z_i, \bar{w} \text{ es un punto regular de } \Delta_\lambda\}.$$

Tenemos que $Z_i = \cup_\lambda Z_{i, \lambda}$.

Vamos a demostrar que $\theta \in \text{Crit}(\mathcal{L}) \iff \bar{u} \in \bar{w} + (\text{Im}(d_\theta \mu))^{\perp_\Sigma}$. Para esto, veremos que es equivalente que θ sea un punto crítico de \mathcal{L} a que $\bar{u} - \bar{w} \in (\text{Im}(d_\theta \mu))^{\perp_\Sigma}$. Denotamos

4. Optimización

$(d_\theta \mu)_{:,j}$ la j -ésima columna de la matriz $d_\theta \mu$. Calculamos el gradiente de $\ell \circ \mu$ en un punto θ . Como $\mathcal{L}(\theta) = \langle \bar{w} - \bar{u}, \bar{w} - \bar{u} \rangle_\Sigma$,

$$\nabla(\mathcal{L})(\theta)_j = 2\langle \bar{w} - \bar{u}, (d_\theta \mu)_{:,j} \rangle.$$

Luego,

$$\begin{aligned} \theta \in \text{Crit}(\mathcal{L}) &\iff \bar{w} - \bar{u} \perp^\Sigma (d_\theta \mu)_{:,j} \forall j = 1 \\ &\iff \bar{w} - \bar{u} \in (\text{Im}(d_\theta \mu))^{\perp \Sigma}. \end{aligned}$$

Por el Lema (4.1.0.2), sabemos que la fibra $\mu^{-1}(\bar{w})$ consiste en una cantidad de finita de clases de equivalencia. Todos los elementos de la clase de $[\theta]$ dan lugar al mismo conjunto $\text{Im}(d_\theta \mu)$. Podemos definir entonces $N_{\Sigma, \bar{w}}$ como la unión de los espacios afines $\bar{w} + (\text{Im}(d_\theta \mu))^{\perp \Sigma}$ para cada clase $[\theta] \in \mu^{-1}(\bar{w})$ tal que algún polinomio $\pi(w^i)$ tiene al menos una raíz múltiple. Como hay finitas clases de equivalencia, se trata de una unión finita.

Luego, $\bar{u} \in \text{pr}_3(Z_{i,\lambda}) \iff \exists \bar{w} \in \Delta_\lambda$ punto regular tal que $\bar{u} \in N_{\Sigma, \bar{w}}$. Como vale que $\pi(w^i)$ tiene alguna raíz múltiple, por la proposición (4.1.0.4), $T_{\bar{w}}\Delta_\lambda \subsetneq \text{Im}(d_\theta \mu)$. Por lo tanto, podemos deducir que $\dim(N_{\Sigma, \bar{w}}) < \dim((T_{\bar{w}}\Delta_\lambda)^{\perp \Sigma}) = k - \dim(\Delta_\lambda)$. Finalmente, $\dim(\text{pr}_3(Z_{i,\lambda})) \leq \dim(\Delta_\lambda) + \max_{\bar{w} \in \text{Reg}(\Delta_\lambda)} \dim(N_{\Sigma, \bar{w}}) < k$. □

A partir del Teorema (4.1.0.2) y del Teorema anterior, podemos entonces deducir que si la función de pérdida es la del error cuadrático, todo punto crítico θ de la función objetivo \mathcal{L} cuya identificación polinomial $\pi(\mu(\theta))$ tenga un *rrmp* compatible con la red se corresponde con un punto crítico de la función de pérdida.

Corolario 4.3.2.1. *Dada una red de profundidad L de arquitectura $(\mathbf{d}, \mathbf{k}, \mathbf{s})$ y el mapa de parametrización μ . Sean k, s el tamaño de filtro y de salto de las matrices convolucionales del espacio de funciones $\mathcal{M} = \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$. Sea $N \in \mathbb{N}$ la cantidad de muestras del conjunto de datos de entrenamiento, sean $X \in \mathbb{R}^{d_0 \times N}, Y \in \mathbb{R}^{d_L \times N}$ datos de entrenamiento genéricos tales que XX^T tiene rango máximo. Sea ℓ una función genérica de error cuadrático como en la Definición 4.3.1.2 y $\mathcal{L} = \ell \circ \mu$.*

*Consideramos un *rrmp* $(\rho|\gamma)$ y λ una partición de $k-1$. Si $(\rho|\gamma)$ es compatible con la red, entonces para todo $\bar{W} \in \mathcal{M}_{\mathbf{d}, \mathbf{k}, \mathbf{s}}$ tal que $\bar{\pi}(\bar{W})$ tiene *rrmp* $(\rho|\gamma)$, es punto regular de Δ_λ y \bar{W} es punto crítico de $\ell|_{\bar{\pi}^{-1}(\Delta_\lambda)}$, existe θ tal que $\mu(\theta) = \bar{W}$ y es un punto crítico de \mathcal{L} .*

4.3.3. Un ejemplo

Daremos un último ejemplo para ilustrar los resultados sobre la optimización de la función de error cuadrático.

4.3. Optimización de la función de error cuadrático

Generamos los datos con el filtro objetivo $u = (2, 0, 5, 0, 2)^T \in \mathbb{R}^5$,

$$\mathcal{D}_N = \{(x_i, y_i) \in \mathbb{R}^5 \times \mathbb{R} : i = 1, \dots, N, y_i = u^T x_i\}$$

y los datos x_i son obtenidos con una distribución Gaussiana estándar.

Estudiaremos el problema de optimización de la función de error cuadrático para distintas redes de dimensión 1 y arquitecturas con dimensiones compatibles. Buscamos que la red aproxime al filtro objetivo u .

Los datos de entrenamiento están normalizados para cumplir con la hipótesis que $XX^T = I$. De esta manera, por el Lema 4.3.1.3, la función objetivo es

$$\mathcal{L}(\theta) = \|\mu(\theta) - u\|^2.$$

Las posibles arquitecturas con salto 1, $d_0 = 5$ y $d_L = 1$ con tamaños de filtro $1 < k_i < 5$ son

$$\mathbf{k} = (4, 2), (3, 3), (3, 2, 2), (2, 2, 2, 2). \quad (4.3.3.1)$$

La única arquitectura *filling* es $\mathbf{k} = (3, 3)$. El filtro objetivo u se identifica con el polinomio

$$\pi(u) = 2x^4 + 5x^2y^2 + 2y^4 = (x^2 + 2y^2)(2x^2 + y^2)$$

que no tiene raíces reales. Por lo que vimos en el Ejemplo 3.2.2.5, $\pi(u) \notin \bar{\pi}(\mathcal{M}_{(4,2)}), \bar{\pi}(\mathcal{M}_{(3,2,2)}), \bar{\pi}(\mathcal{M}_{2,2,2,2})$.

A partir del Teorema 4.1.0.1, sabemos que todo punto crítico θ de \mathcal{L} se corresponde con un punto crítico de ℓ restringida a $\bar{\pi}^{-1}(\Delta_\lambda)$ para una partición λ de 4 que describa la estructura de raíces complejas de $\bar{\pi}(\mu(\theta))$. Para buscar puntos críticos, consideraremos las posibles particiones λ de 4 y computaremos los puntos críticos de $\ell|_{\Delta_\lambda}$.

- $\lambda = (2, 1, 1)$: hay diez puntos críticos complejos sobre Δ_λ , de los cuales 4 son racionales. Son

$$\left(\frac{1}{5}, \pm\frac{9}{5}, \frac{16}{5}, \pm\frac{9}{5}, \frac{1}{5}\right), (0, 0, 5, 0, 2) \text{ y } (2, 0, 5, 0, 0).$$

- $\lambda = (3, 1)$: hay doce puntos críticos complejos sobre Δ_λ , de los cuales cuatro son reales y ninguno es racional.
- $\lambda = (2, 2)$: hay trece puntos críticos complejos sobre Δ_λ , de los cuales cinco son reales y en particular tres son racionales. Son

$$(-1, 0, 2, 0, -1), (0, 0, 5, 0, 0) \text{ y } \left(\frac{7}{3}, 0, \frac{14}{3}, 0, \frac{7}{3}\right).$$

- $\lambda = (4)$: hay diez puntos críticos complejos sobre Δ_λ de los cuales cuatro son reales y racionales. Son

$$(0, 0, 0, 0, 2), (2, 0, 0, 0, 0) \text{ y } \left(\frac{17}{35}, \pm\frac{68}{35}, \frac{102}{35}, \pm\frac{68}{35}, \frac{17}{35}\right)$$

4. Optimización

Para cualquier elección de las arquitecturas listadas en (4.3.3.1), se comprobó en [6] que el algoritmo de descenso del gradiente converge a alguno de los diecisiete puntos críticos reales listados. Más aún, pareciera siempre converger a alguno de los puntos críticos racionales.

Por ejemplo, para la arquitectura con $\mathbf{k} = (4, 2)$, de todas las particiones posibles, la única que da un *rrmp* compatible con la arquitectura es $\lambda = (2, 1, 1)$. Por lo tanto, esperamos que los puntos críticos sean los listados para esta partición. En efecto, el algoritmo de descenso del gradiente converge a $(0, 0, 5, 0, 2)$ o a $(2, 0, 5, 0, 0)$.

En la lista previa se describe a los puntos críticos como filtros de elementos del espacio de funciones pero no se tienen los puntos críticos correspondientes en el espacio de parámetros. Como vimos en el Lema 4.1.0.2, la fibra de cada punto crítico contiene finitas clases de equivalencia. Es posible recuperar los parámetros correspondientes al punto crítico a menos de un factor escalar al iniciar el descenso del gradiente a partir del Corolario 4.2.0.1.

Por ejemplo, fijamos $\mathbf{k} = (4, 3)$ e iniciamos el algoritmo de descenso del gradiente en el punto

$$\theta_0 = ((1, 6, 11, 6), (4, 1)).$$

Esto fija una escritura como producto del polinomio inicial $(x^3 + 6x^2y + 11xy^2 + 6y^3)(4x + y)$. La curva $\mu(\theta_t)$ en el espacio de funciones converge a $\mu(\theta) = (2, 0, 5, 0, 0)$ identificado con el polinomio $p = 2x^4 + 5x^2$. Este polinomio tiene raíces $\{0, 0, i\sqrt{\frac{5}{2}}, -i\sqrt{\frac{5}{2}}\}$ listadas con multiplicidad. En esta arquitectura, solo puede ser parametrizado por filtros de la forma

$$\left(\frac{1}{\kappa}(2, 0, 5, 0, 0), \kappa(1, 0)\right).$$

Con cuentas similares a las realizadas en la demostración del Corolario 4.2.0.1, podemos recuperar $\kappa = \pm\sqrt{\frac{1}{2}\sqrt{31445} - 177}$. Se verificó numéricamente que converge al filtro con $\kappa > 0$.

A. Teorema de Sturm

Las definiciones y los resultados de esta sección pueden hallarse en [14] y en [15].

El teorema de Sturm es un resultado que permite contar la cantidad de raíces reales distintas de un polinomio no nulo en un determinado intervalo. A partir del algoritmo de división de Euclides, se define una sucesión asociada al polinomio estudiado. Existe una versión más general que no utilizaremos en este trabajo.

Definición A.0.0.1. Dados un polinomio $p \in \mathbb{R}[x]$ no nulo, definimos su *sucesión de Sturm* como $f_0(x) = p(x)$, $f_1(x) = p'(x)$, $f_2(x), \dots, f_s(x)$.

A partir de f_2 , cada término f_{i+1} de la sucesión se obtiene cambiando el signo del resto que se calcula al realizar la división euclidiana de los dos términos anteriores f_{i-1} y f_i . La sucesión termina de computarse cuando algún resto es nulo.

Se tiene entonces

$$\begin{aligned} f_0(x) &= q_1(x)f_1(x) - f_2(x) & \deg(f_2(x)) &< \deg(f_1(x)) \\ &\vdots \\ f_{i-1}(x) &= q_i(x)f_i(x) - f_{i+1}(x) & \deg(f_{i+1}(x)) &< \deg(f_i(x)) \\ &\vdots \\ f_{s-1}(x) &= q_s(x)f_s(x) & (f_{s+1}(x) &= 0). \end{aligned}$$

Definimos además el número de variaciones de signos.

Definición A.0.0.2. Dada una sucesión finita de números reales no nulos $\{c_1, c_2, \dots, c_m\}$, el *número de variaciones de signo* cuenta la cantidad de veces que cambia el signo entre términos sucesivos de la sucesión. Es decir, es la cantidad de índices i , $1 \leq i \leq m-1$, que cumplen que $c_i c_{i+1} < 0$.

Por ejemplo, si consideramos la sucesión $\{1, -2, -3, 1, 5\}$, el número de variaciones de signo es 2.

Para un polinomio $p \in \mathbb{R}[x]$ y $f_0(x) = p(x)$, $f_1(x) = p'(x)$, \dots , $f_s(x)$ la sucesión de Sturm asociada, definimos V_a como el número de variaciones de signo de $\{f_0(a), f_1(a), \dots, f_s(a)\}$, la sucesión obtenida al evaluar cada término de la sucesión de Sturm en a .

Teorema A.0.0.1. Sea $p(x)$ un polinomio de grado positivo en $\mathbb{R}[x]$ y $f_0(x) = p(x)$, $f_1(x) = p'(x)$, \dots , f_s la sucesión de Sturm de $p(x)$. Sea $[a, b]$ un intervalo tal que $f(a) \neq 0$ y $f(b) \neq 0$. Entonces, la cantidad de raíces reales distintas en el intervalo (a, b) es $V_a - V_b$.

A. Teorema de Sturm

Es importante observar que el teorema de Sturm no cuenta las raíces reales con multiplicidad sino que indica la cantidad de raíces reales distintas entre sí.

Ejemplo A.0.0.1. Si $p(x) = x^3 + 4x^2 + 5x + 2 = (x + 1)^2(x + 2)$ la sucesión de Sturm es

$$\begin{aligned}f_0(x) &= x^3 + 4x^2 + 5x + 2 \\f_1(x) &= 3x^2 + 8x + 5 \\f_2(x) &= \frac{2}{9}(x + 1).\end{aligned}$$

El polinomio tiene una raíz real doble y una simple en el intervalo $(-3, 0)$.

De $\{f_0(-3), f_1(-3), f_2(-3)\} = \{-4, 8, -\frac{4}{9}\}$ obtenemos que $V_{-3} = 2$.

De $\{f_0(0), f_1(0), f_2(0)\} = \{2, 5, \frac{2}{9}\}$, obtenemos que $V_0 = 0$. Luego, a partir del teorema de Sturm, deducimos que hay dos raíces reales distintas en el intervalo elegido.

A partir del teorema de Sturm se puede contar la cantidad de raíces reales que tiene un polinomio dado en \mathbb{R} . Para eso, se debe tomar el límite de cada término de la sucesión a $+\infty$ y a $-\infty$ y contar el número de variaciones de signo. Definimos $V_{-\infty}$ como el número de variaciones de signo de

$$\left\{ \lim_{x \rightarrow -\infty} f_0(x), \dots, \lim_{x \rightarrow -\infty} f_s(x) \right\}$$

y $V_{+\infty}$ como el número de variaciones de signo de

$$\left\{ \lim_{x \rightarrow +\infty} f_0(x), \dots, \lim_{x \rightarrow +\infty} f_s(x) \right\}.$$

Veamos este procedimiento con el ejemplo anterior.

Ejemplo A.0.0.2. Calculamos los respectivos límites

$$\begin{aligned}\lim_{x \rightarrow -\infty} f_0(x) &= -\infty, & \lim_{x \rightarrow +\infty} f_0(x) &= +\infty, \\ \lim_{x \rightarrow -\infty} f_1(x) &= +\infty, & \lim_{x \rightarrow +\infty} f_1(x) &= +\infty, \\ \lim_{x \rightarrow -\infty} f_2(x) &= -\infty, & \lim_{x \rightarrow +\infty} f_2(x) &= +\infty.\end{aligned}$$

Disponemos en una tabla el signo de cada límite: Observamos que $V_{-\infty} = 2$ y que

	f_0	f_1	f_2
$\lim_{x \rightarrow -\infty} f_i(x)$	-	+	-
$\lim_{x \rightarrow +\infty} f_i(x)$	+	+	+

$V_{+\infty} = 0$ por lo que deducimos que el polinomio tiene dos raíces reales distintas en \mathbb{R} .

Bibliografía

- [1] MathWorks *¿Qué es una red neuronal?* <https://es.mathworks.com/discovery/neural-network.html>
- [2] Goodfellow I., Bengio Y., Courville A. *Deep Learning* (2016) MIT Press
- [3] MathWorks *¿Qué son las redes neuronales convolucionales?* <https://es.mathworks.com/discovery/convolutional-neural-network.html>
- [4] Codificando Bits *La Convolución en las Redes Convolucionales* <https://codificandobits.com/blog/convolucion-redes-convolucionales/>
- [5] Marchetti G. L., Shahverdi V., Mereta S., Trager M., Kohn K. *An Invitation to Neuroalgebraic Geometry*
- [6] Kohn K., Merkh T., Montúfar G., Trager M. *Geometry of linear convolutional networks* (2022) SIAM Journal on Applied Algebra and Geometry, 6(3):368–406
- [7] Kohn K., Montúfar G., Shahverdi V., Trager M. *Function Space and Critical Points of Linear Convolutional Networks* (2024) <https://arxiv.org/abs/2304.05752>
- [8] Kohn K. *The Geometry of the Neuromanifold* <https://www.siam.org/publications/siam-news/articles/the-geometry-of-the-neuromanifold/>
- [9] Hubel D. H., Wiesel T. N. *Receptive fields, binocular interaction and functional architecture in the cat's visual cortex* (1962) J Physiol.
- [10] Fukushima K. *Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position* (1980) Biological Cybernetics
- [11] LeCun Y., Bengio Y., Hinton G. *Deep Learning* (2015) Nature
- [12] Serre T., Poggio T. *A Neuromorphic Approach to Computer Vision* (2010) MIT Open Access Articles
- [13] Distill *Feature Visualization* <https://distill.pub/2017/feature-visualization/>
- [14] Basu S., Pollack R., Roy M. F. *Algorithms in Real Algebraic Geometry* (2006) Springer

Bibliografía

- [15] Gonzalez E., Weinberg D. A. *Root Configurations of Real Univariate Cubics and Quartics* (2018) <https://arxiv.org/abs/1511.07489>
- [16] Arnon D. S. *Geometric reasoning with logic and algebra* (1988) <https://www.sciencedirect.com/science/article/abs/pii/0004370288900495?via%3Dihub>
- [17] Whitney H. *Complex Analytic Varieties* (1972)
- [18] Maple 2021 *Maplesoft, a division of Waterloo Maple Inc., Waterloo, Ontario*