Teorema da Aproximação Universal

Caio Lins

16 de maio de 2021

Conteúdo

1 Introdução

Neste trabalho estudaremos resultados que tratam da tarefa de representar um conjunto de funções por meio de outro, podendo essa representação ser exata ou aproximada. Esse é um tema amplo, presente em diversas áreas da matemática, tanto pura quanto aplicada. Faremos esse estudo em alguns contextos diferentes, sendo o principal deles o da aproximação de funções contínuas por meio de redes neurais.

Começaremos pelo teorema da aproximação de Weierstrass, que estabelece a viabilidade de aproximar arbitrariamente bem funções contínuas em um intervalo da reta por meio de polinômios. Inicialmente provado por Weierstrass em 1885, novas demonstrações surgiram com o tempo. A que apresentaremos — retirada de [You06] — ainda fornecerá uma forma de, dada uma função contínua, obter os polinômios que a aproximam. Essa é uma peculiaridade desse teorema, pois as provas seguintes não são construtivas.

Escolhemos iniciar com esse resultado por sua demonstração ser elementar, não exigindo mais que um curso de análise na reta. Além disso, ele trata de aproximações em um contexto mais simples, com apenas uma variável e no qual as funções aproximadoras, os polinômios, são lineares com relação aos seus coeficientes, os parâmetros da aproximação. Com isso, esperamos acostumar o leitor ao tipo de questão que nos é de interesse.

Em seguida, moveremos nossa atenção para um dos problemas apresentados por David Hilbert no Congresso Internacional de Matemáticos de 1900, mais especificamente, o 13°. Hilbert postulou (utilizando a linguagem matemática de sua época) que existem funções contínuas de $\mathbb{I}^3$ em $\mathbb{R}$, onde $\mathbb{I} = [0, 1]$, que não podem ser expressas por meio da composição e adição de funções de $\mathbb{R}^2$ em $\mathbb{R}$. Décadas após ser postulada, essa conjectura eventualmente foi demonstrada *falsa*. A prova foi dada por Vladimir Igorevich Arnol'd, 14 anos após a morte de Hilbert. Ele e seu orientador de Doutorado, Andrej Nikolajewitsch Kolmogorov, provaram que, na verdade, toda função contínua $f : \mathbb{I}^n \to \mathbb{R}$ pode ser expressa como composições e adições de funções contínuas de $\mathbb{R} \to \mathbb{R}$. A formulação exata que apresentaremos — obtida de [Mor21] — é ainda mais forte que essa.

Naturalmente, a diferença mais óbvia desse resultado para o anterior é que agora nós temos representações *exatas* para nossas funções de interesse, não apenas aproximações. Além disso, esse é um contexto que envolve múltiplas variáveis e no qual as funções aproximadoras não são lineares em seus parâmetros. Portanto, é de se esperar que a demonstração apresentada seja mais complexa. De fato, apesar de não ser extremamente complicada, ela envolve conceitos e resultados sobre espaços métricos aos quais julgamos apropriado dedicar um apêndice.

O último e principal resultado sobre aproximações que apresentaremos é o teorema da aproximação universal, como formulado em [Cyb89]:

Teorema da aproximação universal. *Seja $\sigma : \mathbb{R} \to \mathbb{R}$ uma função contínua discriminatória qualquer. Então, dada qualquer $f : \mathbb{I}^n \to \mathbb{R}$, contínua, e $\varepsilon > 0$, existe uma soma, $G : \mathbb{I}^n \to \mathbb{R}$, da seguinte forma:*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j), \tag{1}$$

onde $\alpha_j, \theta_j \in \mathbb{R}$ *e* $y_j \in \mathbb{R}^n$*, tal que*

$$|G(x) - f(x)| < \varepsilon$$

para todo $x \in \mathbb{I}^n$*.*

Aqui y_j^T é o transposto do vetor y_j, de modo que $y_j^T x$ é o produto interno usual de $\mathbb{R}^n$. A definição de função discriminatória será dada posteriormente.

Antes de apresentar comentários sobre esse teorema, é prudente explicar sua relação com redes neurais. Como descrito em [Lip87], as redes neurais artificiais são algoritmos de computação que surgiram como uma tentativa de espelhar o funcionamento de redes neurais orgânicas, como o cérebro humano. Em essência, seu objetivo é "atingir bom desempenho por meio da densa interconexão de elementos computacionais simples."

Na prática, existem várias formas de atingir esse objetivo. No tipo de rede ao qual daremos enfoque, o processo de computação é realizado por um conjuntos de nós, organizados em camadas ordenadas, em que a camada inicial é composta por n nós de `input` e a camada final, por m nós de `output`. Entre eles a rede pode possuir camadas intermediárias de tamanho variado. Dados n valores reais para os nós de `input`, a rede produz, nos nós de `output`, m valores, também reais. Logo, ela pode ser considerada uma função de $\mathbb{R}^n$ em $\mathbb{R}^m$.

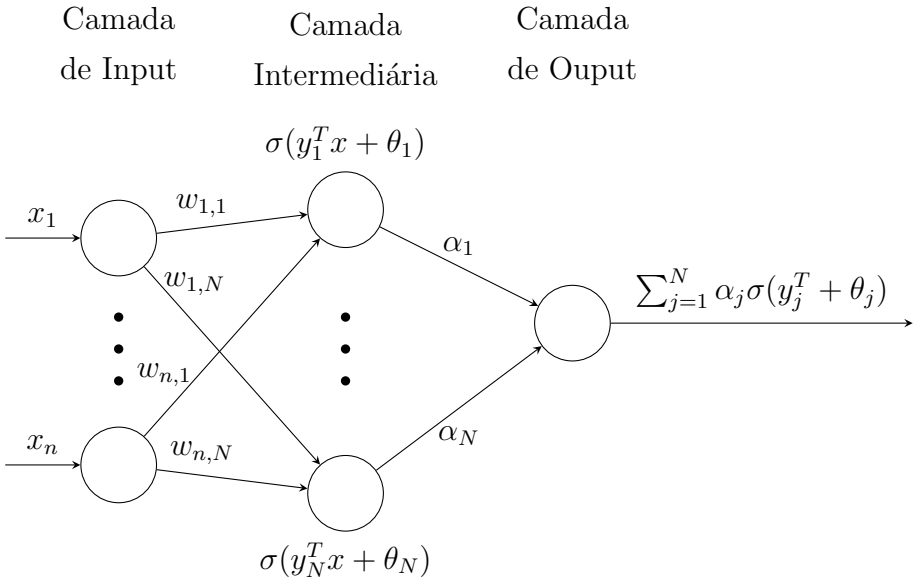

Figura 1: Rede neural com apenas uma camada intermediária. Aqui temos $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^T$ e $y_j^T x = \begin{bmatrix} w_{1,j} & \cdots & w_{n,j} \end{bmatrix}$, o vetor dos pesos de cada nó intermediário. Na última camada ocorre apenas uma combinação linear.

A partir da segunda camada, cada nó está conectado a todos os nós da camada anterior por meio de uma aresta, a qual possui um peso. Esse nó computa uma combinação linear, utilizando como coeficientes os pesos das arestas, dos valores armazenados pelos nós da camada anterior, soma um fator de correção ao valor obtido e passa o resultado por uma função não-linear, obtendo assim um

valor próprio. O caso em que $m = 1$ e há apenas uma camada intermediária de N nós, ao qual nos atentaremos, está representado na figura 1, onde σ é a não-linearidade.

Repare que o output da rede é exatamente a expressão em (1). Ou seja, o teorema da aproximação universal, doravante conhecido como **TAU**, estabelece a possibilidade de aproximar arbitrariamente bem funções reais contínuas em $\mathbb{I}^n$ por meio de redes neurais com uma camada intermediária. Essa é uma pergunta de extrema relevância, tanto teórica como prática pois, como apontado em [Lip87], redes neurais artificiais possuem diversas aplicações em campos voltados ao desenvolvimento de classificadores robustos, como a teoria de reconhecimento de fala e de imagens. Saber que é possível realizar aproximações arbitrariamente boas utilizando redes neurais artificiais dá mais segurança e incentivo à pesquisa nessas áreas.

A demonstração do **TAU** envolve resultados de teoria da medida e análise funcional, os quais são importantes por si mesmos e, por isso, são discutidos em seções próprias. Alguns conhecimentos necessários para entendê-los são apresentados como outro apêndice. De fato, a demonstração desse resultado é mais densa que a dos outros, apesar de que ele, ao contrário do problema de Hilbert, lida apenas com aproximações. Uma possível explicação para essa diferença será apresentada na seção correspondente.

2 Teorema da aproximação de Weierstrass

Como dito na introdução, desejamos mostrar que dada uma função contínua $f : [a, b] \to \mathbb{R}$, podemos aproximá-la arbitrariamente bem por funções polinomiais $p : [a, b] \to \mathbb{R}$.

Em outras palavras, seja $C([a, b])$ o espaço vetorial das funções contínuas em $[a, b]$. Indicamos por $\|\varphi\|_\infty$ a norma do supremo de uma função limitada $\varphi : [a, b] \to \mathbb{R}$, ou seja,

$$\|\varphi\|_\infty = \sup \{|\varphi(x)| : x \in [a, b]\} .$$

Então é verdade que

Teorema 2.1. *Dada $f \in C([a, b])$, para todo $\varepsilon > 0$ existe um polinômio $p : [a, b] \to \mathbb{R}$ tal que*

$$\|f - p\|_\infty < \varepsilon.$$

Inicialmente, observamos que basta provar o teorema para o caso $f \in C([0, 1])$. De fato, dada $f \in C([a, b])$, considere o homeomorfismo $\varphi : [0, 1] \to [a, b]$ dado por $\varphi(x) = a + (b - a)x$, cuja inversa é $\varphi^{-1} : [a, b] \to [0, 1]$ dada por $\varphi^{-1}(x) = \frac{x-a}{b-a}$. Então a função $g = f \circ \varphi$ pertence a $C([0, 1])$ e, dado $\varepsilon > 0$, se existe um polinômio $p(x)$ com $\|g - p\|_\infty < \varepsilon$, temos também, como φ^{-1} é um polinômio de grau 1,

$$\left\|g \circ \varphi^{-1} - p \circ \varphi^{-1}\right\|_\infty < \varepsilon.$$

Como $g \circ \varphi^{-1} = f$ e $p \circ \varphi^{-1}$ é um polinômio, o resultado vale também para $C([a, b])$.

Em seguida, devemos definir a classe de polinômios que utilizaremos na demonstração.

Definição 2.1. Dada $g : \mathbb{R} \to \mathbb{R}$ definimos o n-ésimo polinômio de Bernstein de g como

$$B_n(x, g) := \sum_{k=0}^{n} g\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}. \tag{2}$$

Note a semelhança entre os polinômios de Bernstein e a expansão binomial de $(1 + (1 - x))^n$. De fato, temos $B_n(x, 1) = (1 + (1 - x))^n = 1$. Mais geralmente, para toda constante $c \in \mathbb{R}$ tem-se $B_n(x, c) = c$.

Utilizaremos essa semelhança para obter algumas identidades essenciais para a demonstração do teorema 2.1. Dados p e q reais, começamos considerando a expansão binomial de $(p + q)^n$:

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}.$$

Considerando ambos lados da igualdade como funções de p, podemos derivá-los com relação a essa variável, obtendo

$$n(p+q)^{n-1} = \sum_{k=0}^{n} k \binom{n}{k} p^{k-1} q^{n-k}.$$

Multiplicando ambos lados por p/n, ficamos com

$$p(p+q)^{n-1} = \sum_{k=0}^{n} \frac{k}{n} \binom{n}{k} p^k q^{n-k}. \tag{3}$$

Essa é a primeira identidade, válida para todos $p, q \in \mathbb{R}$. Derivando novamente com relação a p e multiplicando ambos lados por p/n obtemos

$$p^2 \left(1 - \frac{1}{n}\right)(p+q)^{n-2} + \frac{p}{n}(p+q)^{n-1} = \sum_{k=0}^{n} \frac{k^2}{n^2} \binom{n}{k} p^k q^{n-k}, \tag{4}$$

a segunda identidade que utilizaremos.

Como consideramos $f, g \in C([0, 1])$, segue da Definição 2.1 que se $f \geq 0$, então $B_n(x, f) \geq 0$ e, se $f \leq g$, então $B_n(x, f) \leq B_n(x, g)$.

Com essas ferramentas, podemos então apresentar a

Demonstração do teorema 2.1. Observamos inicialmente que como f é uma função contínua definida em um compacto, é uniformemente contínua. Portanto, dado $\varepsilon > 0$, existe $\delta > 0$ tal que, se $x, y \in [0, 1]$ satisfazem $|x - y| < \delta$, então

$$|f(x) - f(y)| < \frac{\varepsilon}{2}.$$

Agora, definimos $M := \|f\|_\infty$ e fixamos $\xi \in [0, 1]$. Logo, se $|x - \xi| \geq \delta$ temos

$$|f(x) - f(\xi)| \leq 2M \leq 2M \left(\frac{x-\xi}{\delta}\right)^2.$$

Combinando as duas últimas desigualdades, concluímos que para todo $x \in [0, 1]$ vale

$$|f(x) - f(\xi)| \leq 2M \left(\frac{x-\xi}{\delta}\right)^2 + \frac{\varepsilon}{2}. \tag{5}$$

Vamos aproximar f pelos seus polinômios de Bernstein. Seja $B_n(x,f)$ o n-ésimo polinômio de Bernstein de f, avaliado em x. Então

$$\begin{aligned}
|B_n(x,f) - f(\xi)| &= |B_n(x, f - f(\xi))| \\
&\leq B_n\left(x, 2M\left(\frac{x-\xi}{\delta}\right)^2 + \frac{\varepsilon}{2}\right) \\
&= \frac{2M}{\delta^2} B_n(x,(x-\xi)^2) + \frac{\varepsilon}{2} \\
&+ \frac{2M}{\delta^2}\left(B_n(x,x^2) + B_n(x,-2x\xi) + \xi^2\right) + \frac{\varepsilon}{2} \\
&= \frac{2M}{\delta^2}\left(B_n(x,x^2) - 2\xi B_n(x,x) + \xi^2\right) + \frac{\varepsilon}{2}. \qquad (6)
\end{aligned}$$

Aqui fizemos uso das propriedades de $B_n(x,f)$ que seguem de $x \in [0,1]$, discutidas anteriormente. Utilizando as equações (3) e (4), com a substituição $p = x$ e $q = 1 - x$, concluímos que

$$B_n(x,x) = x.$$

e que

$$B_n(x,x^2) = x^2\left(1 - \frac{1}{n}\right) + \frac{x}{n}.$$

Substituindo em (6), ficamos com

$$\begin{aligned}
\frac{2M}{\delta^2}\left(B_n(x,x^2) - 2\xi B_n(x,x) + \xi^2\right) + \frac{\varepsilon}{2} &= \frac{2M}{\delta^2}\left(x^2\left(1-\frac{1}{n}\right) + \frac{x}{n} - 2\xi x + \xi^2\right) + \frac{\varepsilon}{2} \\
&= \frac{2M}{\delta^2}\left(x^2 + \frac{x-x^2}{n} - 2\xi x + \xi^2\right) + \frac{\varepsilon}{2} \\
&= \frac{\varepsilon}{2} + \frac{2M}{n\delta^2}(x-x^2) + \frac{2M}{\delta^2}(x-\xi)^2.
\end{aligned}$$

Sendo assim,

$$|B_n(x,f) - f(\xi)| \leq \frac{\varepsilon}{2} + \frac{2M}{n\delta^2}(x-x^2) + \frac{2M}{\delta^2}(x-\xi)^2.$$

Como essa desigualdade vale para todo $x \in [0,1]$, em especial é válida para $x = \xi$. Fazendo essa substituição, obtemos

$$|B_n(\xi,f) - f(\xi)| \leq \frac{\varepsilon}{2} + \frac{2M}{n\delta^2}(\xi - \xi^2).$$

Facilmente podemos verificar que $\xi - \xi^2 \leq \frac{1}{4}$ para todo $\xi \in [0,1]$. Logo,

$$|B_n(\xi,f) - f(\xi)| \leq \frac{\varepsilon}{2} + \frac{M}{2n\delta^2}.$$

Por fim, tomando $n > \frac{M}{\varepsilon\delta^2}$, temos $\frac{M}{2n\delta^2} < \frac{\varepsilon}{2}$ e, assim,

$$|B_n(\xi,f) - f(\xi)| < \varepsilon.$$

Como o valor de n obtido para que essa desigualdade seja satisfeita depende apenas de ε (lembramos que δ depende apenas de ε, pela continuidade uniforme de f), ela é válida para todo $\xi \in [0,1]$, ou seja,

$$\|B_n(\cdot,f) - f\|_\infty < \varepsilon. \qquad \square$$

3 Décimo Terceiro Problema de Hilbert

Aqui provaremos a possibilidade de representar qualquer função contínua $f : \mathbb{I}^n \to \mathbb{R}$ por meio de composições e adições de funções contínuas de $\mathbb{R} \to \mathbb{R}$. Com o trabalho de vários matemáticos, esse resultado foi generalizado. Uma dessas generalizações é apresentada no teorema a seguir.

O leitor interassado poderá encontrar mais informações em [Mor21].

Teorema 3.1 (Kolmogorov, Arnol'd, Kahane, Lorentz e Sprecher)**.** *Para todo $n \in \mathbb{N}$ com $n \geq 2$, existem números reais $\lambda_1, \lambda_2, \ldots, \lambda_n$ e funções contínuas $\varphi_k : \mathbb{I} \to \mathbb{R}$, para $k = 1, \ldots, 2n+1$, com a propriedade de que para toda função contínua $f : \mathbb{I}^n \to \mathbb{R}$ existe uma função contínua $g : \mathbb{R} \to \mathbb{R}$ tal que, para todo $(x_1, \ldots, x_n) \in \mathbb{I}^n$,*

$$f(x_1, \ldots, x_n) = \sum_{k=1}^{2n+1} g(\lambda_1 \varphi_k(x_1) + \cdots + \lambda_n \varphi_k(x_n)). \tag{7}$$

Observação. Denotamos o espaço das funções contínuas de $\mathbb{I}^n$ em $\mathbb{R}$ por $C(\mathbb{I}^n)$. Notamos que esse espaço, com a norma do supremo, se torna um espaço métrico completo, com a distância dada por

$$d\left((f_1, \ldots, f_n), (g_1, \ldots, g_n)\right) = \max\left\{\|f_1 - g_1\|_\infty, \ldots, \|f_n - g_n\|_\infty\right\}. \tag{8}$$

O leitor pode encontrar resultados e definições elementares relativas a espaços métricos (inclusive o teorema da categoria de Baire, que será usado a seguir) no apêndice A.

Aqui nos atentaremos ao caso especial em que $n = 2$:

Teorema 3.2. *Existem $\lambda \in \mathbb{R}$ e funções $(\varphi_1, \ldots, \varphi_5) \in [C(\mathbb{I})]^5$ tais que, para toda função $f \in C(\mathbb{I}^2)$ existe $g : \mathbb{R} \to \mathbb{R}$, contínua, satisfazendo, para todos $(x_1, x_2) \in \mathbb{I}^2$,*

$$f(x_1, x_2) = \sum_{k=1}^{5} g(\varphi_k(x_1) + \lambda \varphi_k(x_2)).$$

Observação. Por uma questão de economia de notação, definimos

$$\Phi_k(x, y) = \sum_{k=1}^{5} \varphi_k(x) + \lambda \varphi_k(y).$$

Para a prova que apresentaremos, necessitamos de alguns lemas. O primeiro deles clarifica a existência e a escolha de λ.

Lema 3.1. *Existe um número real λ tal que, para quaisquer $x_1, x_2, x_3, x_4 \in \mathbb{Q}$,*

$$x_1 + \lambda y_1 = x_2 + \lambda y_2 \text{ implica } x_1 = x_2 \text{ e } y_1 = y_2.$$

Demonstração. Basta escolher $\lambda \in \mathbb{R}\backslash\mathbb{Q}$, pois se $x_1 + \lambda y_1 = x_2 + \lambda x_2$, então $x_1 - x_2 = \lambda(y_2 - y_1)$. Como $x_1 - x_2 \in \mathbb{Q}$ necessariametente o lado direito deve ser 0, ou seja, $y_2 = y_1$ e $x_2 = x_1$. □

Lema 3.2. *Fixe λ satisfazendo o lema 3.1. Seja $f \in C(\mathbb{I}^2)$ com $\|f\|_\infty = 1$. Seja U_f o suconjunto de $[C(\mathbb{I})]^5$ tal que $(\varphi_1, \ldots, \varphi_5) \in U_f$ se, e somente se, existe uma $g \in C(\mathbb{R})$ tal que*

$$|g(t)| \leq \frac{1}{7} \text{ para todo } t \in \mathbb{R}, \tag{9}$$

e

$$\left| f(x,y) - \sum_{i=1}^{5} (g \circ \Phi_k)(x,y)) \right| < \frac{7}{8}, \text{ para todo } x, y \in \mathbb{I}^2. \tag{10}$$

Então U_f é um suconjunto aberto e denso de $[C(\mathbb{I})]^5$.

Enquanto é fácil perceber que U_f é aberto, pois se $\varphi = (\varphi_1, \ldots, \varphi_5)$ satisfaz (9) e (10) para uma dada g, então todas as tuplas suficientemente perto de φ também satisfarão, a prova da densidade é mais extensa e será omitida, por não ser do interesse desse trabalho. Nos restringimos a mencionar que ela necessita do lema 3.1.

Lema 3.3. *Seja λ como no lema 3.1. Então existem $\varphi_1, \ldots, \varphi_5 \in C(\mathbb{I})$ com a propriedade de que, dada $f \in C(\mathbb{I}^2)$, existe uma $g \in C(\mathbb{R})$ satisfazendo*

$$|g(t)| \leq \frac{1}{7}\|f\|_\infty \text{ para todo } t \in \mathbb{R},$$

e

$$\left\| f - \sum_{k=1}^{5} g \circ \Phi_k \right\|_\infty < \frac{8}{9}\|f\|_\infty.$$

Demonstração. Podemos supor, sem perda de generalidade, que $\|f\|_\infty = 1$. Seja $(h_j)_{j\in\mathbb{N}}$ uma sequência de funções pertencentes a $C(\mathbb{I}^2)$ tal que o conjunto $\{h_j : j \in \mathbb{N}\}$ é denso na esfera unitária de $C(\mathbb{I}^2)$. Na notação do lema 3.2, cada h_j determina um conjunto $U_j = U_{h_j} \subset [C(\mathbb{I})]^5$ aberto e denso. Pelo teorema da categoria de Baire, o conjunto

$$V = \bigcap_{j\in\mathbb{N}} U_j$$

é denso em $[C(\mathbb{I})]^5$. Pela densidade de $\{h_j : j \in \mathbb{N}\}$, existe $m \in \mathbb{N}$ tal que $\|f - h_m\|_\infty < 1/72$. Além disso, pelo lema 3.2 podemos tomar $(\varphi_1, \ldots, \varphi_5) \in V \subset U_m$ tal que exista $g \in C(\mathbb{I})$ satisfazendo

$$|g(t)| \leq \frac{1}{7} \text{ para todo } t \in \mathbb{R},$$

e

$$\left\| h_m - \sum_{k=1}^{5} g \circ \Phi_k \right\|_\infty < \frac{7}{8}.$$

Com isso,

$$\left\| f - \sum_{k=1}^{5} g \circ \Phi_k \right\|_\infty \leq \|f - h_m\|_\infty + \left\| h_m - \sum_{k=1}^{5} g \circ \Phi_k \right\|_\infty < \frac{1}{72} + \frac{7}{8} = \frac{8}{9},$$

o que termina a prova. □

Agora podemos enunciar a demonstração do teorema 3.2.

Demonstração do teorema 3.2. Pelo lema 3.3, podemos fixar $\lambda \in \mathbb{R}$ e $\varphi_1, \ldots, \varphi_5 \in C(\mathbb{I})$ tais que, dada $f \in C(\mathbb{I}^2)$, existe $g_0 \in C(\mathbb{R})$ satisfazendo

$$|g_0(t)| \leq \frac{1}{7}\|f\|_\infty \text{ para todo } t \in \mathbb{R}, \tag{11}$$

e

$$\left\| f - \sum_{k=1}^{5} g_0 \circ \Phi_k \right\|_\infty < \frac{8}{9}\|f\|_\infty. \tag{12}$$

Defina $f_0 := f$. Então, supondo definidas $f_0, \ldots, f_n$, $g_0, \ldots, g_n$, ponha $f_{n+1} = f_n - \sum_{k=1}^{5} g_n \circ \Phi_k$. Logo, existe g_{n+1} satisfazendo (11) e (12) com g_{n+1} no lugar de g_0 e f_{n+1} no lugar de f.

Dessa forma, temos $\|g_n\|_\infty \leq \frac{1}{7}\|f_n\|_\infty$ e $\|f_{n+1}\|_\infty = \left\|f_n - \sum_{k=1}^{5} g_n \circ \Phi_k\right\|_\infty \leq \frac{8}{9}\|f_n\|_\infty$. Portanto,

$$\|f_n\|_\infty \leq \left(\frac{8}{9}\right)^n \|f_0\|_\infty = \left(\frac{8}{9}\right)^n \|f\|_\infty,$$

e

$$\|g_n\|_\infty < \frac{1}{7}\left(\frac{8}{9}\right)^n \|f\|.$$

Sendo assim, a série $\sum g_n$ converge em módulo para uma certa $g \in C(\mathbb{R})$ e, com isso,

$$f = \sum_{n\in\mathbb{N}} f_n - f_{n+1} = \sum_{n\in\mathbb{N}} \sum_{k=1}^{5} g_n \circ \Phi_k = \sum_{k=1}^{n} \left(\sum_{n\in\mathbb{N}} g_n\right) \circ \Phi_k = \sum_{k=1}^{5} g \circ \Phi_k,$$

o que conclui a demonstração. □

4 Teorema de Hahn-Banach

Um dos dois teoremas essenciais para a prova do teorema da aproximação universal é o teorema de Hahn-Banach. Em análise funcional ele é relevante por garantir a existência de funcionais lineares limitados em um espaço vetorial normado qualquer. Nossa exposição é inteiramente baseada em [Oli12] e o leitor que quiser revisar os conceitos de análise funcional necessários para compreendê-la poderá encontrá-los no apêndice B.

Definição 4.1. Se V é um espaço vetorial real, um *funcional sublinear* em V é uma função $p : V \to \mathbb{R}$ tal que

$$p(x + y) \leq p(x) + p(y) \text{ e } p(\lambda x) = \lambda p(x) \text{ para todos } x, y \in V \text{ e } \lambda \in \mathbb{R}_{\geq 0}.$$

Observe que nada se exige sobre o sinal de p. Um exemplo de funcional sublinear é uma seminorma em V.

Teorema 4.1 (Hahn-Banach)**.** *Sejam V um espaço vetorial real, M um subespaço de V e p um funcional sublinear em V. Se f é um funcional linear em M dominado por p, ou seja, tal que $f(v) \leq p(v)$ para todo $v \in M$, então existe um funcional linear F em V, que coincide com f em M e que também é dominado por p. O funcional F é dito* extensão de Hahn-Banach *de f.*

Em outras palavras, se, em um espaço vetorial, temos um funcional sublinear que domina um funcional linear definido em um subespaço, podemos extender esse funcional ao espaço todo, mantendo a relação de dominância.

Antes de apresentar a demonstração do teorema 4.1, vamos introduzir alguns conceitos de Teoria dos Conjuntos.

4.1 Lema de Zorn

Dados conjuntos X uma *relação* de X em Y é um subconjunto R de $X \times Y$. Escreveremos xRy para indicar que $(x, y) \in R$. Perceba que uma função $f : X \to Y$ é uma relação de X em Y dada por $(x, y) \in f$ se, e somente se, $y = f(x)$. Quando $R \subset X \times X$, diremos que R é uma relação em X.

Uma *relação de ordem parcial* em X é uma relação que satisfaz, dados $x, y, z \in X$:

i) xRx (*reflexividade*);

ii) Se xRy e yRx, então $x = y$ (*antisimetria*);

iii) Se xRy e yRz, então xRz (*transitividade*).

O termo *parcial* é utilizado para indicar que podem existir elementos de X não relacionados por essa ordem. Caso tenhamos xRy ou yRx para todos $x, y \in X$, então R é dita uma relação de ordem *total*. Utilizaremos o símbolo $\prec$ para indicar uma ordem parcial. Dizemos então que $(X, \prec)$ é um conjunto parcialmente ordenado. Se $\prec$ também é total, $(X, \prec)$ é totalmente ordenado. Claramente todo subconjunto de X também é parcialmente ordenado, com a ordem induzida por $\prec$.

Um *elemento maximal* de um conjunto ordenado X é um elemento x tal que se $y \in X$ com $x \prec y$, então $y = x$. Uma *cota superior* de um subconjunto $Y \subset X$ é um elemento $x \in X$ tal que $y \prec x$ para todo $y \in Y$. Observe que se $\prec$ não é total, não necessariamente um elemento maximal de $Y \subset X$ é uma cota superior de Y.

Dois exemplos usuais de conjuntos ordenados são $\mathbb{R}$ e $\mathcal{P}(X)$, para um dado conjunto X. O primeiro é um conjunto totalmente ordenado pela ordem usual $\leq$, o qual não possui elemento maximal. O segundo se torna um conjunto ordenado ao dizermos que dados $A, B \subset X$, temos $A \prec B$ se, e somente se $A \subset B$. Essa ordem não é total e o único elemento maximal de $\mathcal{P}(X)$ é o próprio X.

Axioma (lema de Zorn)**.** *Todo conjunto parcialmente ordenado, tal que todos seus subconjuntos totalmente ordenados possuam cota superior, possui elemento maximal.*

Encerramos essa apresentação com um resultado que exemplifica como o lema de Zorn geralmente é utilizado.

Proposição 4.1. *Todo espaço vetorial não trivial, ou seja, que possui elementos não nulos, possui uma base.*

Demonstração. Seja V o espaço vetorial em questão, e A a coleção de todos os subconjuntos linearmente independentes de V. Claramente essa coleção é parcialmente ordenada pela relação de

inclusão. Dada uma subcoleção $B \subset A$ totalmente ordenada, considere o conjunto

$$C = \bigcup_{\beta \in B} \beta.$$

Afirmamos que $C \in A$. De fato, dado um conjunto finito $\{x_1, \ldots, x_n\} \subset C$, sejam $\beta_1, \ldots, \beta_n$ os conjuntos de B tais que $x_i \in \beta_i$ para todo $i \leq n$. Como B é totalmente ordenado, existe β_k tal que $\beta_i \subset \beta_k$ para todo $i \leq n$. Logo $\{x_i, \ldots, x_n\} \subset \beta_k$ e, como β_k é linearmente independente, também o são os x_i. Dessa forma, C é um conjunto linearmente independente, pertencente a A e que claramente é uma cota superior de B.

Portanto, aplicando o lema de Zorn obtemos um elemento maximal Λ de A. Denotando por $\operatorname{Lin}\Lambda$ o conjunto formado pelas combinações lineares de elementos de Λ, afirmamos que $\operatorname{Lin}\Lambda = A$. Com efeito, supondo, por absurdo, que exista $v \in V \setminus \operatorname{Lin}\Lambda$, temos que $\Lambda \cup \{v\}$ é linearmente independente, ou seja, pertence a A e contém Λ, um contradição pois Λ é maximal. □

4.2 Demonstração do teorema de Hahn-Banach

Seja $\{g_\lambda\}_{\lambda \in L}$, para algum conjunto de índices L, a coleção das extensões lineares de f em conjuntos M_λ tais que $M \subset M_\lambda \subset V$, que são dominadas por p. Como $f \in \{g_\lambda\}_{\lambda \in L}$, essa coleção é não-vazia e, portanto, podemos ordená-la parcialmente utilizando a relação de inclusão, considerando que g_λ é um subconjunto de $M_\lambda \times \mathbb{R}$. Mais especificamente, diremos que $g_\alpha \subset g_\beta$ se tivermos $M_\alpha \subset M_\beta$ e $g_\beta \mid_{M_\alpha} = g_\alpha$, ou seja, se g_β for uma extensão de g_α. Desejamos mostrar que existe um elemento maximal de $\{g_\lambda\}_{\lambda \in L}$. Dado $L' \subset L$ tal que $\{g_\lambda\}_{\lambda \in L'}$ é uma subcoleção totalmente ordenada, definimos g como

$$g := \bigcup_{\lambda \in L'} g_\lambda,$$

ou seja, g é um funcional linear em $\bigcup_{\lambda \in L} M_\lambda$ tal que $g(v) = g_\lambda(v)$, onde $v \in M_\lambda$. De fato g está bem definida, pois se $v \in M_{\lambda_1} \cap M_{\lambda_2}$ com $\lambda_1 \neq \lambda_2$, ambos pertencentes a L', então podemos supor, sem perda de generalidade, que $g_{\lambda_2} \subset g_{\lambda_1}$ e, assim, $g_{\lambda_1}(v) = g_{\lambda_2}(v)$, para $v \in M_{\lambda_1} \cap M_{\lambda_2} = M_{\lambda_2}$. É fácil verificar que g estende f e é dominado por p. Logo, g é uma cota superior de $\{g_\lambda\}_{\lambda \in L'}$ e, pelo lema de Zorn, existe $F \in \{g_\lambda\}_{\lambda \in L}$, definido em $W \subset V$, maximal.

Intuitivamente é claro que F deve estar definida em todo o espaço V, de modo que é a extensão de Hahn-Banach de f. Para demonstrar esse fato, suponha, por absurdo, que ele seja falso, ou seja, que exista $\eta \in V \setminus W$. Nossa estratégia será construir um uma extensão G de F, dominada por p, definida em $U = \operatorname{Lin}(W \cup \{\eta\})$, contradizendo a maximalidade de F.

Para definirmos G, basta atribuirmos um valor para $G(\eta)$. De fato, todo elemento de U é da forma $w + \alpha\eta$, onde $w \in W$ e $\alpha \in \mathbb{R}$. Logo, pondo

$$G(w + \alpha\eta) = F(w) + \alpha G(\eta),$$

claramente G é uma extensão de F. O passo crucial é definir $G(\eta)$ de forma que G seja dominada por p. Para tanto, vamos nos atentar a algumas desigualdades. Dados $w_1, w_2 \in W$, vale

$$F(w_1) + F(w_2) = F(w_1 + w_2) \leq p(w_1 + w_2) \leq p(w_1 - \eta) + p(\eta + w_2).$$

Equivalentemente:

$$F(w_1) - p(w_1 - \eta) \leq p(\eta + w_2) - F(w_2).$$

Ou seja, temos

$$\sup\{F(w_1) - p(w_1 - \eta) : w_1 \in W\} \leq \inf\{p(\eta + w_2) - F(w_2) : w_2 \in W\}.$$

Defina, então, $G(\eta)$ de modo que

$$\sup\{F(w_1) - p(w_1 - \eta) : w_1 \in W\} \leq G(\eta) \leq \inf\{p(\eta + w_2) - F(w_2) : w_2 \in W\}.$$

Dessa forma, dado $w + \alpha\eta \in U$, se $\alpha > 0$ temos

$$\begin{aligned} G(w + \alpha\eta) &= F(w) + \alpha G(\eta) \\ &\leq F(w) + \alpha\left(p\left(\frac{w}{\alpha} + \eta\right) - F\left(\frac{w}{\alpha}\right)\right) \\ &= F(w) + p(w + \alpha\eta) - F(w) \\ &= p(w + \alpha\eta). \end{aligned}$$

Caso $\alpha < 0$:

$$\begin{aligned} G(w + \alpha\eta) &= F(w) + \alpha G(\eta) \\ &\leq F(w) + \alpha\left(F\left(\frac{w}{|\alpha|}\right) - p\left(\frac{w}{|\alpha|} - \eta\right)\right) \\ &= F(w) - F(w) - p(-w - \alpha\eta) \\ &= p(w + \alpha\eta). \end{aligned}$$

Se $\alpha = 0$, G coincide com F e claramente é dominada por p. Sendo assim, chegamos a uma contradição com a maximalidade de F, o que conclui a prova. □

No caso especial em que p é uma seminorma, a condição $f \leq p$ é equivalente a $|f| \leq p$, pois $-f(x) = f(-x) \leq p(-x) = p(x)$.

4.3 Aplicações do teorema de Hahn-Banach

Aqui vamos utilizar o teorema de Hahn-Banach para provar um resultado sobre subespaços vetoriais densos que será utilizado na prova do teorema da aproximação universal. Começamos apresentando uma aplicação do teorema de Hahn-Banach para ilustrar como ele é geralmente utilizado.

Teorema 4.2. *Seja V um espaço vetorial normado. Então:*

i) Se $0 \neq v \in V$, então existe $f \in V^*$ tal que $f(v) = \|v\|$ e $\|f\| = 1$;

ii) Se v e w são elementos distintos de V, então existe $f \in V^*$ tal que $f(v) \neq f(w)$;

iii) Se $v \in V$ satisfaz $f(v) = 0$ para todo $f \in V^*$, então $v = 0$;

iv) Se $v \in V$, então

$$\|v\| = \sup_{0 \neq f \in V^*} \frac{|f(v)|}{\|f\|} = \max_{0 \neq f \in V^*} \frac{|f(v)|}{\|f\|}.$$

Demonstração. Para provar *i*, basta aplicar Hahn-Banach com o funcional sublinear dado pela norma, o subespaço $M = \text{Lin}(\{v\})$ e o funcional $f : M \to \mathbb{R}$ dado por $f(\alpha v) = \alpha\|v\|$. Sendo F a extensão de Hahn-Banach de f, temos $F(v) = \|v\|$ e, como o funcional sublinear é a norma de V, vale $|F(x)| \leq \|x\|$ para todo $x \in V$. Como há um caso de igualdade, vale $\|F\| = 1$. Em *ii*, como $v \neq w$, então $v - w \neq 0$. Por *i*, isso implica a existência de um funcional f tal que $f(v - w) = \|v - w\| \neq 0$, ou seja, $f(v) \neq f(w)$. Claramente *iii* segue diretamente de *ii*. Como, para todo $f \in V^*$, vale $|f(v)| \leq \|f\|\|v\|$, temos

$$\|v\| \geq \sup_{0 \neq f \in V^*} \frac{|f(v)|}{\|f\|}.$$

Por *i*, existe $g \in V^*$ tal que $g(v) = \|v\|$ e $\|g\| = 1$, de modo que

$$\|v\| = \frac{|g(v)|}{\|g\|},$$

ou seja, vale a igualdade e podemos trocar sup por max. □

A próxima proposição será útil para provar o principal resultado dessa seção.

Proposição 4.2. *Seja V um espaço vetorial normado e M um subespaço próprio fechado de V. Dado $v \in V \backslash M$, definimos $\delta = d(v, M) := \inf\{\|v - w\| : w \in M\}$. Então existe um funcional $f \in V^*$ tal que $\|f\| = 1$, $f(v) = \delta$ e $f\,|_M = 0$.*

Demonstração. Como M é fechado, $\delta > 0$. Defina o funcional $g : \text{Lin}(\{v\} \cup M) \to \mathbb{R}$ dado por

$$g(\alpha v + w) = \alpha\delta,$$

onde $\alpha \in \mathbb{R}$ e $w \in M$. Claramente temos $g\,|_M = 0$ e $g(v) = \delta$. Agora perceba que

$$|g(\alpha v + w)| = |\alpha|\delta \leq |\alpha|\|v + w/\alpha\| = \|\alpha v + w\|,$$

de modo que $\|g\| \leq 1$. Além disso, para todo $w \in M$ vale

$$\|g\| \geq \frac{|g(v + w)|}{\|v + w\|} = \frac{\delta}{\|v + w\|}.$$

Portanto,

$$\|g\| \geq \sup_{w \in M} \frac{\delta}{\|v + w\|} = \frac{\delta}{\inf_{w \in M}\|v - w\|} = 1.$$

Logo, $\|g\| \geq 1$ e, com isso, $\|g\| = 1$. Agora podemos aplicar Hahn-Banach, obtendo a extensão f de g, utilizando como funcional sublinear a norma em V. □

Agora o teorema que será usada na prova do teorema da Aproximação Universal.

Teorema 4.3. *Um subespaço M de um espaço vetorial V é denso se, e somente se, o único elemento de V^* que se anula em M é o funcional nulo.*

Demonstração. Suponha que N é um subespaço denso de V e seja $f \in V^*$ um funcional que se anula em M. Dado $v \in V \backslash M$, seja $(x_n)_{n\in\mathbb{N}}$ tal que $x_n \in M$ para todo n e $x_n \to v$. Como f é uma transformação linear limitada, é contínua e, portanto, $f(x_n) \to f(v)$. Porém $f(x_n) = 0$ para todo n, ou seja, $f(x_n) \to 0 = f(v)$, de modo que $f = 0$.

Reciprocamente, se o único elemento de V^* que se anula em M é o funcional nulo, então necessariamente V é denso. Caso contrário, $\overline{M}$ seria um subespaço próprio fechado de V e, pela proposição 4.2 existiria um funcional não nulo que se anula em $\overline{M}$ e, consequentemente, em M. □

5 Teorema da Representação de Riesz

Esse teorema possui diversas formulações. Em geral, elas têm como objetivo representar os funcionais lineares contínuos em um dado espaço vetorial de uma maneira mais natural, associando-os a elementos daquele mesmo espaço ou de outro.

Apesar de a versão que abordaremos inicialmente não ser a utilizada na demonstração do teorema da aproximação universal, acreditamos que é importante apresentar esse teorema primeiro no contexto natural de espaços de Hilbert.

Novamente, nossa exposição desse resultado clássico da análise funcional foi formulada com base em [Oli12] e o leitor que quiser recordar conceitos fundamentais poderá encontrálos no apêndice B.

5.1 Em espaços de Hilbert

Da forma que será primeiramente enunciado, esse resultado trata da associação natural que existe entre um espaço de Hilbert H e o seu dual, H^*. Apesar de ser um fato de certa forma trivial para espaços de dimensão finita, sua demonstração não é tão óbvia para espaços de Hilbert em geral. Para demonstrá-lo, precisamos, antes, de passar por três resultados preliminares. O leitor poderá encontrar alguns dos conceitos de análise funcional utilizados no apêndice correspondente.

Lema 5.1. *Dados v, w pertencentes a $(V, \langle\cdot,\cdot\rangle)$, tem-se que $v \perp w$, ou seja, $\langle v, w\rangle = 0$, se, e somente se,*

$$\|v + \lambda w\| \geq \|v\|. \tag{13}$$

para todo $\lambda \in \mathbb{K}$.

Demonstração. Evidentemente temos

$$0 \leq \|v + \lambda w\|^2 = \|v\|^2 + 2\mathrm{Re}(\langle v, w\rangle) + |\lambda|^2\|w\|^2.$$

Se $v \perp w$, então temos

$$\begin{aligned}\|v + \lambda w\|^2 &= \|v\|^2 + |\lambda|^2\|w\|^2 \\ &\geq \|v\|^2,\end{aligned}$$

de onde a desigualdade (13) segue. Reciprocamente, se vale (13) para todo $\lambda \in \mathbb{K}$, em especial tomando $\lambda = -\langle w, v\rangle/\|w\|^2$ e elevando ambos lados ao quadrado ficamos com $0 \leq -|\langle v, w\rangle|^2$, o que implica $v \perp w$. □

Lema 5.2 (Lei do paralelogramo)**.** *Dados v, w pertencentes a $(V, \langle\cdot,\cdot\rangle)$, tem-se*

$$\|v+w\|^2 + \|v-w\|^2 = 2\|v\|^2 + 2\|w\|^2.$$

Como a demonstração desse lema consiste simplesmente em expandir o lado esquerdo da igualdade e usar as propridedades de produto interno, será deixada a cargo do leitor.

Antes do próximo resultado, uma definição. Dado um subespaço E de um espaço com produto interno V, definimos

$$E^\perp = \{v \in V : \langle v, w\rangle = 0 \text{ para todo } w \in E\}.$$

Teorema 5.1 (Projeção ortogonal)**.** *Se E é subespaço vetorial fechado de um espaço de Hilbert H, então*

$$H = E \oplus E^\perp.$$

Demonstração. Dado $v \in H$, definimos $\delta = \inf\{\|v-w\| : w \in E\}$. Seja $(w_n)_{n\in\mathbb{N}}$ uma sequência de elementos de E tais que $\|v - w_n\| \to \delta$. Então, sendo k e ℓ números naturais, aplicando a lei do paralelogramo para os vetores $w_k - v$ e $w_\ell - v$ obtemos:

$$2\|w_k - v\|^2 + 2\|w_\ell - v\|^2 = \|w_k + w_\ell - 2v\|^2 + \|w_k - w_\ell\|,$$

o que implica, remanejando e lembrando que $(w_k + w_\ell)/2 \in E$,

$$\begin{aligned}\|w_k - w_\ell\| &= 2\|w_k - v\|^2 + 2\|w_\ell - v\|^2 - 4\|(w_k + w_\ell)/2 - v\|^2 \\ &\leq 2\|w_k - v\|^2 + 2\|w_\ell - v\|^2 - 4\delta^2.\end{aligned}$$

Com isso, concluímos que (w_n) é uma sequência de Cauchy. Como H é um espaço de Hilbert, temos que $w_n \to w \in E$, pois E é fechado e, pela continuidade da norma, temos $\|v - w\| = \delta$.

Intuitivamente, w é o elemento de E mais próximo de v. Logo, é razoável esperar que ele seja a projeção ortogonal de v em E. Para confirmar essa suspeita, devemos verificar que $v - w \in E^\perp$. De fato, para todo $\lambda \in \mathbb{K}$ e todo $u \in E$ temos

$$\|(v-w) + \lambda u\| = \|v + (-w + \lambda u)\| \geq \delta = \|v - w\|.$$

Portanto, pelo lema 5.1, concluímos que $v - w \in E^\perp$. Sendo assim, temos $v = w + (v-w)$, onde $w \in E$ e $v - w \in E^\perp$. Para mostrar a unicidade dessa decomposição, suponha que $v = w_1 + u_1 = w_2 + u_2$, onde $w_1, w_2 \in E$ e $u_1, u_2 \in E^\perp$. Então

$$w_1 - w_2 = u_2 - u_1 \in E \cap E^\perp,$$

o que implica $w_1 - w_2 = u_2 - u_1 = 0$, ou seja, $w_1 = w_2$ e $u_1 = u_2$. □

Agora podemos prosseguir para o principal resultado dessa subseção.

Teorema 5.2 (Teorema da representação de Riesz em espaços de Hilbert)**.** *Dado um espaço de Hilbert H e seu dual H^*, a função*

$$\begin{aligned}\gamma : &H \to H^* \\ &v \mapsto \gamma(v) = f_v,\end{aligned}$$

tal que $f_v = \langle v, \cdot \rangle$, é uma isometria antilinear e sobrejetiva em H^*.

Demonstração. Para mostrar que γ é uma isometria, provaremos que $\|f_v\| = \|v\|$ para todo $v \in H$. De fato, se $v = 0$ isso é evidente. Fixado $v \in H \setminus \{0\}$, pela desigualdade de Cauchy-Schwarz temos $|f_v(w)| = |\langle v, w \rangle| \leq \|v\|\|w\|$. Logo, $\|f_v\| \leq \|v\|$. Por outro lado, temos $\|v\|^2 = \langle v, v \rangle = f_v(v) \leq \|f_v\|\|v\|$, ou seja, $\|v\| \leq \|f_v\|$ e acabamos.

Além disso claramente γ é antilinear, pois

$$\gamma(\alpha v + w) = \langle \alpha v + w, \cdot \rangle = \bar{\alpha}\langle v, \cdot \rangle + \langle w, \cdot \rangle = \bar{\alpha}\gamma(v) + \gamma(w).$$

A parte menos óbvia, e que dá nome ao teorema, é a da sobrejetividade. Repare que, com essa propriedade, todo funcional $f \in H^*$ fica unicamente associado a um determinado $v \in H$ tal que $f = \gamma(v)$. Dizemos, então, que v *representa* f. Para demonstrá-la, utilizaremos o teorema 5.1.

Dado $f \in H^*$, se $f = 0$ então naturalmente $f = \gamma(0)$. Se $f \neq 0$, repare que $\ker(f)$, o núcleo de f, é um subespaço próprio fechado de H. Pelo teorema 5.1, temos

$$H = \ker(f) \oplus \ker(f)^\perp,$$

de onde concluímos que existe $w \in \ker(f)^\perp$ satisfazendo $\|w\| = 1$. Agora uma observação simples, porém fundamental: para todos $u, v \in H$ temos $f(v)u - f(u)v \in \ker(f)$. Logo, para todo $v \in H$ temos

$$\langle w, f(v)w - f(w)v \rangle = 0,$$

o que implica, desenvolvendo,

$$f(v) = \langle \overline{f(w)}w, v \rangle.$$

Sendo assim, $f = \gamma(\overline{f(w)}w)$. □

6 Teorema da Aproximação Universal

Nesta seção apresentaremos o principal resultado estudado: o teorema da aproximação universal, como formulado em [Cyb89]. Como o leitor pode lembrar, quando abordamos o 13° problema de Hilbert, enunciamos o caso geral e provamos um caso particular do teorema da superposição de Kolmogorov, segundo o qual podemos representar de forma *exata* funções contínuas de $\mathbb{I}^n$ em $\mathbb{R}$ utilizando a superposição e composição de funções contínuas de $\mathbb{I}$ em $\mathbb{R}$, de uma maneira parecida como a apresentada em (1). Entretanto, como comentado em [Cyb89], a diferença crucial entre esses dois resultados é que no teorema de Kolmogorov nos é permitido usar uma classe muito maior de não-linearidades, as quais variam de acordo com a função a ser aproximada. Isso fornece uma intuição do porquê obter representações exatas é uma tarefa intratável no caso de redes neurais.

6.1 O Teorema

Denotaremos o espaço das funções reais contínuas em $\mathbb{I}^n$ por $C(\mathbb{I}^n)$ e o espaço das medidas de Borel finitas, com sinal e regulares em $\mathbb{I}^n$ por $M(\mathbb{I}^n)$. Queremos entender sob quais condições as somas da forma

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j)$$

são densas em $C(\mathbb{I}^n)$ com respeito à norma do supremo. Faremos isso primeiro para uma classe mais geral de funções σ e depois mostraremos que as sigmoides, ou seja, funções σ tais que

$$\sigma(x) \to \begin{cases} 1, \text{ se } x \to +\infty \\ 0, \text{ se } x \to -\infty \end{cases},$$

pertencem a essa classe. Essa classe de funções é importante pois as funções de ativação de redes neurais (as não-linearidades em cada nó) são, em geral, sigmoides.

Definição 6.1. Dizemos que σ é *discriminatória* se, para uma medida $\mu \in M(I^n)$, termos

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

para todos $y \in \mathbb{R}^n$ e $\theta \in \mathbb{R}$ implica em $\mu = 0$.

Teorema da aproximação universal. *Seja σ uma função contínua discriminatória qualquer. Então as somas finitas da forma*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j). \tag{14}$$

são densas em $C(\mathbb{I}^n)$. Em outras palavras, dada qualquer $f \in C(\mathbb{I}^n)$ e $\varepsilon > 0$, existe uma soma, $G(x)$, da forma acima, tal que

$$\|G - f\|_\infty < \varepsilon.$$

Demonstração. Seja $S \subset C(\mathbb{I}^n)$ o conjunto das funções descritas por (14). Claramente ele é um subespaço de $C(\mathbb{I}^n)$, o qual nós afirmamos ser denso. Suponha, por absurdo, que ele não o seja, isto é, se R é o fecho de S, suponha que tenhamos $C(\mathbb{I}^n)\backslash \neq \emptyset$. Então, pelo teorema 4.3, existe um funcional linear limitado $L : C(\mathbb{I}^n) \to \mathbb{R}$, não nulo, tal que $L(R) = 0$. Pelo teorema da representação de Riesz, esse funcional é da forma

$$L(h) = \int_{\mathbb{I}^n} h(x) \, \mathrm{d}\mu(x),$$

para alguma $\mu \in M(\mathbb{I}^n)$ e toda $h \in C(\mathbb{I}^n)$. Em particular, como $\sigma(y^T x + \theta) \in S$ para todos $y \in \mathbb{R}^n$ e $\theta \in \mathbb{R}$, temos

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

para todos y, θ. Porém, como σ é, por hipótese, discriminatória, isso implica L ser o funcional nulo, onde chegamos a uma contradição. Portanto, $R = C(\mathbb{I}^n)$, ou seja, S é denso em $C(\mathbb{I}^n)$. □

Tendo provado o teorema para funções discriminatórias, nos resta mostrar que, de fato, sigmoides são discriminatórias.

Lema 6.1. *Todas sigmoides limitadas e mensuráveis são discriminatórias.*

Demonstração. Seja σ uma sigmoide e suponha que, para uma dada medida $\mu \in M(\mathbb{I}^n)$, tenhamos

$$\int_{\mathbb{I}^n} \sigma(y^T x + \theta) \, \mathrm{d}\mu(x)$$

para todos $y \in \mathbb{R}^n$ e $\theta \in \mathbb{R}$. Desejamos provar que isso implica $\mu = 0$.

Começamos a demonstração reparando que, para quaisquer x, y, θ, φ temos

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

Logo, a família de funções $\sigma_\lambda(x) = \sigma(\lambda(y^T x + \theta) + \varphi)$ é limitada e converge pontualmente à função

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

quando $\lambda \to +\infty$.

Seja $\Pi_{y,\theta}$ o hiperplano definido por $\left\{x \in \mathbb{R}^n : y^T x + \theta = 0\right\}$ e seja $H_{y,\theta}$ o meio-espaço aberto definido por $\left\{x \in \mathbb{R} : y^T x + \theta > 0\right\}$. Então, pelo teorema da convergência dominada de Lebesgue, temos

$$\begin{aligned} 0 &= \int_{\mathbb{I}^n} \sigma_\lambda(x) \, \mathrm{d}\mu(x) \\ &= \int_{\mathbb{I}^n} \gamma(x) \, \mathrm{d}\mu(x) \\ &= \sigma(\varphi)\mu(\Pi_{y,\theta}) + \mu(H_{y,\theta}) \end{aligned}$$

para todos φ, θ e y.

Agora mostraremos que se □

Referências

[Lip87] R. Lippmann. «An introduction to computing with neural nets». Em: *IEEE ASSP Magazine* 4.2 (1987), pp. 4–22. DOI: 10.1109/MASSP.1987.1165576.

[Cyb89] George Cybenko. «Approximation by superpositions of a sigmoidal function». Em: *Math. Control. Signals Syst.* 2 (1989), pp. 303–314. DOI: 10.1007/BF02551274. URL: https://doi.org/10.1007/BF02551274.

[You06] Matt Young. *The Stone-Weierstrass theorem*. Jan. de 2006.

[Oli12] César R. de Oliveira. *Introdução à análise funcional*. Ed. por IMPA. IMPA, 2012.

[Mor21] Sidney A. Morris. «Hilbert 13: are there any genuine continuous multivariate real-valued functions?» Em: *Bulletin (New Series) of the American Mathematical Society* 58.1 (jan. de 2021), pp. 107–118. URL: https://doi.org/10.1090/bull/1698.

A Elementos de Espaços Métricos

Aqui apresentamos noções básicas relativas a espaços métricos, amplamente utilizadas ao longo do texto.

Definição A.1. Dado um conjunto X qualquer, uma *métrica* em X é uma função $d : X \times X \to \mathbb{R}$ tal que:

i) $d(x, x) = 0$;

ii) $d(x, y) > 0$ se $x \neq y$;

iii) $d(x, y) = d(y, x)$;

iv) $d(x, z) \leq d(x, y) + d(y, z)$.

Definição A.2. Um *espaço métrico* é um par (X, d) onde X é um conjunto e d é uma métrica em X.

Por vezes, onde não houver prejuízo ao entendimento do texto, utilizaremos apenas o nome do conjunto para nos referirmos ao espaço métrico por ele formado.

Definição A.3. Um suconjunto M de um espaço métrico X é dito *limitado* se existe $c \in \mathbb{R}$ tal que $d(x, y) \leq c$ para todos $x, y \in M$. Nesse caso, o definimimos o *diâmetro* de M, denotado por $\operatorname{diam} M$, como $\sup \{d(x, y) : x, y \in M\}$. Se M é ilimitado, ou seja, para todo $c > 0$ existem $x, y \in M$ com $d(x, y) > c$, dizemos que $\operatorname{diam} M = \infty$.

Definição A.4. Dado um espaço métrico X e um ponto $a \in X$, chamamos de *bola aberta de raio r centrada em a* o conjunto

$$B(a, r) := \{x \in X : d(x, a) < r\}.$$

Definição A.5. Dado um espaço métrico X e um subconjunto $Y \subset X$, chamamos de *interior* de Y, e denotamos por $\operatorname{int} Y$, o subconjunto de Y formado pelos elementos $a \in Y$ tais que existe $r > 0$ satisfazendo $B(a, r) \subset Y$.

Definição A.6. Um subconjunto A de um espaço métrico X é dito *aberto* se $A = \operatorname{int} A$.

Definição A.7. Dado um subconjunto M de um espaço métrico X, um ponto $x \in X$ é dito *aderente* a M se toda bola aberta centrada em x tiver interseção não-vazia com M. Chamamos de *fecho* de M, e denotamos por $\overline{M}$, o conjunto dos pontos de aderência de M.

Definição A.8. Um subconjunto F de um espaço métrico X é dito *fechado* se $F = \overline{F}$.

Definição A.9. Dada uma sequência $(x_n)_{n \in \mathbb{N}}$ de elementos do espaço métrico X, dizemos que $(x_n)_{n \in \mathbb{N}}$ *converge* para $L \in X$ se, dado $\varepsilon > 0$, existe $n_0 \in \mathbb{N}$ tal que, para $n \geq n_0$, vale $d(x_n, L) < \varepsilon$. Se para todo $L \in X$ é falso que $\lim x_n = L$, dizemos que (x_n) é *divergente*.

Definição A.10. Uma sequência $(x_n)_{n\in\mathbb{N}}$ de elementos do espaço métrico X é dita *de Cauchy* se, dado $\varepsilon > 0$, existe $n_0 \in \mathbb{N}$ tal que, para $n, m \geq n_0$, vale $d(x_n, x_m) < \varepsilon$. Equivalentemente, (x_n) é de Cauchy se, para $n \geq n_0$, vale $d(x_n, x_{n+p}) < \varepsilon$ para todo $p \in \mathbb{N}$.

Proposição A.1. *Toda sequência convergente $(x_n)_{n\in\mathbb{N}}$ no espaço métrico X é de Cauchy.*

Demonstração. Seja $L = \lim x_n$. Dado $\varepsilon > 0$, tome $n_0 \in \mathbb{N}$ de modo que, para $n \geq n_0$, valha $d(x_n, L) < \varepsilon/2$. Então, se $n, m \geq n_0$ temos

$$d(x_n, x_m) \leq d(x_n, L) + d(x_m, L) = \varepsilon/2 + \varepsilon/2 = \varepsilon. \qquad \square$$

Exemplo 1. Embora toda sequência convergente seja de Cauchy, é falso que dado um espaço métrico qualquer, toda sequência de Cauchy convirja para um ponto pertencente a ele. Por exemplo, considerando o conjunto $\mathbb{Q}$ com a métrica d induzida pela métrica de $\mathbb{R}$, temos que toda sequência de racionais convergindo para um irracional é de Cauchy, mas diverge em $\mathbb{Q}$.

Proposição A.2. *Se $(x_n)_{n\in\mathbb{N}}$ em um espaço métrico X é de Cauchy e possui um valor de aderência (ou seja, existe uma subsequência convergente (x_{n_k}) de (x_n)), então (x_n) converge para esse valor de aderência.*

Demonstração. Seja $L \in X$ o limite da subsequência (x_{n_k}). Então, dado $\varepsilon > 0$ conseguimos obter $k_0 \in \mathbb{N}$ tal que, se $k > k_0$, então $d(x_{n_k}, L) < \varepsilon/2$. Também conseguimos $n_0 \in \mathbb{N}$ tal que, se $n, m \geq n_0$ então $d(x_n, x_m) < \varepsilon/2$. Tome $\ell > \max\{n_0, k_0\}$. Então claramente $n_\ell \geq \ell$ e, com isso,

$$d(x_\ell, L) \leq d(x_\ell, x_{n_\ell}) + d(x_{n_\ell}, L) < \varepsilon/2 + \varepsilon/2 = \varepsilon. \qquad \square$$

Definição A.11. Um espaço métrico X é dito *completo* se toda sequência de Cauchy em X converge para um elemento de X.

Proposição A.3. *Um espaço métrico X é completo se, e somente se, dada uma sequência decrescente $F_1 \supset F_2 \supset \cdots$ de conjuntos não-vazios fechados em X, tais que $\lim \operatorname{diam} F_n = 0$, existe $a \in X$ com*

$$\{a\} = \bigcap_{n=1}^{\infty} F_i.$$

Demonstração. Suponha que X seja completo e considere $(F_n)_{n\in\mathbb{N}}$ como no enunciado do teorema. Para cada conjunto F_n, escolha $x_n \in F_n$, formando uma sequência $(x_n)_{n\in\mathbb{N}}$ de Cauchy. De fato, como $\lim \operatorname{diam} F_n = 0$, dado $\varepsilon > 0$ existe $n_0 \in \mathbb{N}$ tal que para $n \geq n_0$, temos $d(x, y) < \varepsilon$ para todos $x, y \in F_n$. De $F_1 \supset F_2 \supset \cdots$ concluímos que $n, m > n_0$ implicam $x_n, x_m \in F_{n_0}$ o que implica $d(x_n, x_m) < \varepsilon$.

Da completude de X concluímos que existe $a = \lim x_n$. Como todos F_n são fechados e, para $m \geq n$ temos $x_m \in F_n$, conclui-se que $a \in F_n$ para todo $n \in \mathbb{N}$, ou seja,

$$a \in \bigcap_{n=1}^{\infty} F_n.$$

Suponha, agora, que X seja um espaço métrico no qual toda sequência de fechados como a do enunciado convirja. Seja $(x_n)_{n\in\mathbb{N}}$ uma sequência de Cauchy em X. Defina, para cada $n \in \mathbb{N}$, o conjunto $F_n = \{x_n, x_{n+1}, \dots\}$. Então $(\overline{F_n})_{n\in\mathbb{N}}$ é uma sequência decrescente de conjuntos fechados tais que $\lim \operatorname{diam} \overline{F_n} = \lim \operatorname{diam} F_n = 0$. Por hipótese, existe $a \in \bigcap \overline{F_n}$. Como a é limite de sequência de pontos de F_k para todo $k \in \mathbb{N}$, para cada k podemos escolher $a_{n_k} \in F_k$ de modo que $d(a, a_{n_k}) < 1/k$ e, assim, $\lim a_{n_k} = a$. Claramente $n_k > k$ para todo $k \in \mathbb{N}$, portanto, passando a uma subsequência se necessário, a_{n_k} é subsequência de (x_n) o que implica, como (x_n) é de Cauchy, $\lim x_n = a$. □

Teorema A.1 (Teorema da Categoria de Baire)**.** *Se X é um espaço métrico completo e $A_1, A_2, \dots$ são abertos densos em X, então*

$$A = \bigcap_{i=1}^{\infty} A_i$$

é denso em X.

Demonstração. Devemos mostrar que dado V um conjunto aberto em X, temos $A \cap V \neq \emptyset$. Nossa estratégia será construir uma sequência decrescente $F_1 \supset F_2 \supset \cdots$ de conjuntos fechados não-vazios tais que $\lim \operatorname{diam} F_n = 0$ e, para todo $n \in \mathbb{N}$, $F_n \subset A_n \cap V$. Então, pela Proposição A.3, o ponto x que satisfaz $\{x\} = \bigcap F_n$ é tal que $x \in V$ e $x \in F_n \subset A_n$ para todo n, ou seja, $x \in A$ e, portanto, $x \in A \cap V$.

Começamos obeservando que, como A_1 é denso, $A_1 \cap V$ é um conjunto aberto não-vazio. Logo, existe B_1 bola aberta não-vazia de raio menor que 1, tal que $\overline{B_1} \subset A_1 \cap V$. Suponha, agora, definidos $B_1, \dots, B_n$ de forma que, para todo $1 < k \leq n$, B_k é uma bola aberta não-vazia de raio menor que $1/k$ tal que $\overline{B_k} \subset V \cap A_k \cap B_{k-1}$. Novamente, como A_{n+1} é denso, $A_{n+1} \cap B_n$ é um conjunto aberto não vazio. Logo, definimos B_{n+1} como uma bola aberta não-vazia contida em $A_{n+1} \cap B_n$, de raio menor que $1/(n+1)$ tal que $\overline{B_{n+1}} \subset A_{n+1} \cap B_n \subset A_{n+1} \cap V$.

Com isso, obtemos uma sequência decrescente $B_1 \supset \cdots \supset B_n \supset \cdots$ de bolas abertas não-vazias, com o raio de B_n menor que $1/n$, cujos fechos $\overline{B_1} \supset \cdots \supset \overline{B_n} \supset \cdots$ formam uma sequência decrescente de conjuntos fechados não-vazios, com $\operatorname{diam} \overline{B_n} \leq 1/n$ e $\overline{B_n} \subset A_n \cap V$ para todo $n \in \mathbb{N}$, o que, como apontado anteriormente, termina a prova. □

Terminaremos essa seção com a definição de função contínua, que será usada posteriormente.

Definição A.12. Dados espaços métricos (X, d_X) e (Y, d_Y), uma função $f : X \to Y$ é dita *contínua* em $a \in X$ se, para todo $\varepsilon > 0$, existe $\delta > 0$ tal que, se $d_X(x, a) < \delta$, então $d_Y(f(x), f(a)) < \varepsilon$.

B Elementos de Análise Funcional

Aqui apresentaremos noções elementares que mesclam conceitos de Análise e de Álgebra Linear.

Definição B.1. Dado um espaço vetorial V sobre um corpo $\mathbb{K}$, uma *norma* em V é uma função $\|\cdot\| : V \to [0, +\infty)$ tal que

i) $\|v\| = 0$ se, e somente se $v = 0$ (normas são *positivas definidas*);

ii) $\|\alpha v\| = |\alpha|\|v\|$ para todo $v \in V$ e $\alpha \in \mathbb{K}$ (*homogeneidade absoluta*);

iii) $\|v + w\| \leq \|v\| + \|w\|$ para todos $v, w \in V$ (*desigualdade triangular*).

Uma função que cumpre as propriedades *ii* e *iii* acima é denominada uma *semi-norma*. Um espaço vetorial junto de sua norma e denominado um *espaço vetorial normado*. A menos de onde houver ambiguidade, utilizaremos o mesmo símbolo $\|\cdot\|$ para nos referir a normas de qualquer espaço vetorial normado. O vetor dentro do símbolo deixará claro a norma de qual espaço está sendo utilizada.

É interessante perceber que todo espaço vetorial normado é um espaço métrico, quando se introduz nele a métrica induzida pela norma: $(v, w) \mapsto \|v - w\|$. As propriedades de uma métrica são facilmente verificadas.

Definição B.2. Uma transformação linear $T : V \to W$ entre dois espaços vetoriais normados é dita *limitada* se existe $C \in [0, +\infty)$ tal que $\|Tv\| \leq C\|v\|$ para todo $v \in V$.

Essa definição naturalmente é diferente da definição usual de função limitada, pois, como $T(\lambda v) = \lambda(Tv)$, impossível termos $\|Tv\| \leq C$ para todo $v \in V$ se T é não nula.

O próximo teorema ilustra a utilidade dessa definição de transformação limitada.

Teorema B.1. *Seja $T : V \to W$ uma transformação linear entre espaços vetoriais normados. As seguintes afirmativas são equivalentes:*

i) T é contínua;

ii) T é contínua em 0;

iii) T é limitada.

Demonstração. A implição de *i* para *ii* é óbvia. Suponha, agora, que T é contínua em 0. Tomando $\varepsilon = 1$, conseguimos $\delta > 0$ tal que se $\|v\| \leq \delta$, então $\|Tv\| < 1$. Agora, dado $v \in V$, $v \neq 0$, tome o vetor $v' := \delta v/\|v\|$. Claramente temos $\|v'\| = \delta$, ou seja, $\|Tv'\| < 1$, o que implica $\|Tv\| < \delta^{-1}\|v\|$. Logo, tomando $C = \delta^{-1}$, T é limitada. Por fim, se T é limitada, dados $a \in V$ e ε, tome $\delta < \varepsilon/C$. Com isso se $\|x - a\| < \delta$ temos:

$$\|Tx - Ta\| = \|T(x - a)\| \leq C\|x - a\| < C \cdot \varepsilon/C = \varepsilon.$$

Portanto, T é contínua. □

Denotamos pro $L(V, W)$ o conjunto das transformações lineares limitadas de V em W. Esse conjunto, com as operações usuais de adição e multiplicação por escalar, é um espaço vetorial. Podemos, ainda, definir nele uma norma, dada por

$$\|T\| := \sup\left\{\frac{\|Tv\|}{\|v\|} : v \in V \setminus \{0\}\right\} \tag{15}$$

$$= \sup\{\|Tv\| : v \in V, \|v\| = 1\} \tag{16}$$

$$= \inf\{C : \|Tv\| \leq C\|v\| \text{ para todo } v \in V\}. \tag{17}$$

Claramente os dois primeiros conjuntos são iguais, logo seus supremos também o são. Para ver que vale (17), perceba que, por (15), temos $\|Tv\| \leq \|T\|\|v\|$ para todo $v \in V$. Portanto, sendo A o conjunto em (17), temos $\|T\| \in A$, ou seja, $\inf A \leq \|T\|$. Além disso, dado $C \in A$, temos $C \geq \|Tv\|/\|v\|$ para todo $v \in V - \{0\}$. Logo, $\|T\| \leq C$ e, com isso, $\|T\| \leq \inf A$. Sendo assim, vale (17). As propriedades i), ii) e iii) decorrem diretamente das propriedades das normas de V e W e são facilmente verificadas. Sempre trataremos o conjunto $L(V, W)$ como um espaço vetorial normado munido dessa norma, a qual é denominada *norma de operador*. No caso especial em que $W = \mathbb{K}$, o espaço $L(V, \mathbb{K})$ é denominado *espaço dual* de V e é denotado por V^*. Os elementos de V^* são denominados *funcionais lineares*.

Definição B.3. Um *produto interno* no espaço vetorial V é um funcional $(v, w) \mapsto \langle v, w\rangle$, de $V \times V$ em $\mathbb{K}$, de modo que, para todos $u, v, w \in V$ e $\alpha \in \mathbb{K}$ valha

i) $\langle \alpha v + u, w\rangle = \bar{\alpha}\langle v, w\rangle + \langle u, w\rangle$;

ii) $\langle v, w\rangle = \overline{\langle v, w\rangle}$;

iii) $\langle v, v\rangle \geq$ e $\langle v, v\rangle = 0$ se, e somente se, $v = 0$.

O par $(V, \langle\cdot,\cdot\rangle)$ é chamado, então, de *espaço com produto interno* ou *espaço pré-Hilbertiano*.

Como de costume, $\bar{\alpha}$ denota o complexo conjugado de α. Em um espaço vetorial com produto interno, a menos que se dia o contrário, tem-se $\|v\| = \sqrt{\langle v, v\rangle}$. De fato essa função define uma norma em V, como será mostrado a seguir. Antes, porém, devemos estabelecer um resultado fundamental em espaços com produto interno.

Teorema B.2 (Desigualdade de Cauchy-Schwarz)**.** *Em um espaço com produto interno* $(V, \langle\cdot,\cdot\rangle)$*, tem-se*

$$|\langle v, w\rangle| \leq \|v\|\|w\|$$

para todos $v, w \in V$*, com a igualdade ocorrendo se, e somente se,* $v = \lambda w$*.*

Demonstração. Se $\langle v, w\rangle = 0$, a igualdade é imediata. Caso contrário, então $w \neq 0$ e, para $\lambda \in \mathbb{K}$ tem-se

$$0 \leq \langle v - \lambda w, v - \lambda w\rangle = \|v\|^2 - \lambda\langle v, w\rangle - \bar{\lambda}\langle w, v\rangle + |\lambda|^2\|w\|^2.$$

tomando $\lambda = \langle w, v\rangle/\|w\|^2$ vem

$$0 \leq \|v\|^2 - |\langle v, w\rangle|^2/\|w\|^2,$$

o que implica

$$|\langle v, w\rangle|^2 \leq \|v\|^2\|w\|^2,$$

de onde o resultado se segue. Naturalmente a igualdade ocorre se e somente se existe λ tal que $\langle v - \lambda w, v - \lambda w\rangle = 0$, ou seja, $v = \lambda w$. □

É fácil ver que as duas primeiras propriedades da norma são satisfeitas por $\|\cdot\|$ como definida anteriormente. Para verficar a desigualdade triangular perceba que

$$\begin{aligned}\|v+w\|^2 &= \|v\|^2 + 2\mathrm{Re}\langle v, w\rangle + \|w\|^2 \\ &\leq \|v\|^2 + 2|\langle v, w\rangle| + \|w\|^2 \\ &\leq \|v\|^2 + 2\|v\|\|w\| + \|w\|^2 \\ &= (\|v\| + \|w\|)^2.\end{aligned}$$

Portanto, $\|\cdot\|$ de fato é uma norma, denominada *norma induzida pelo produto interno*.

Definição B.4. Um *espaço de Hilbert* é um espaço com produto interno completo com relação à norma induzida pelo produto interno.