

USANDO PROBABILIDADES PARA APROXIMAR FUNÇÕES POR POLINÓMIOS

JOEL MOREIRA

RESUMO. Uma ideia central em Análise moderna é a de aproximar objectos potencialmente mal comportados por objectos mais simples. O Teorema de Aproximação de Weierstrass é um teorema clássico em análise sobre aproximação de funções contínuas por polinómios. Neste artigo apresentamos uma demonstração desse teorema, proposta por Bernstein, ilustrando como a Teoria das Probabilidades permite obter demonstrações mais intuitivas em análise.

1. INTRODUÇÃO

Seja $f : [0, 1] \rightarrow \mathbb{R}$ uma função contínua. A função f pode ser muito mal comportada; por exemplo, pode não ter derivada em ponto nenhum. Assim, para lidar com funções contínuas genéricas é muito útil poder aproximá-las por funções mais regulares. E não podíamos pedir melhor que polinómios para fazer este trabalho. Com esta ideia, Weierstrass, em 1885, demonstrou o seguinte resultado:

TEOREMA DE WEIERSTRASS. Seja $f : [0, 1] \rightarrow \mathbb{R}$ uma função contínua e $\varepsilon > 0$ um número real. Então existe um polinómio $p(x)$ de coeficientes reais tal que, para cada $x \in [0, 1]$, temos $|f(x) - p(x)| < \varepsilon$.

Este teorema é fundamental porque permite que muitas propriedades sobre funções contínuas sejam apenas verificadas para polinómios. Vejamos uma aplicação deste teorema.

Seja $f : [0, 1] \rightarrow \mathbb{R}$ uma função contínua. Vamos provar que

$$(1) \quad \lim_{n \rightarrow \infty} n \int_0^1 x^{n+2} f(x) dx = f(1).$$

Isto verifica-se imediatamente para monómios, calculando directamente o integral. Como o lado esquerdo é linear em f , temos que a propriedade é satisfeita para qualquer polinómio. Finalmente, para uma função contínua arbitrária, aplicando o Teorema de Weierstrass encontramos, para cada $\varepsilon > 0$, um polinómio g tal que $r(x) = g(x) - f(x)$ satisfaz $|r(x)| < \varepsilon$ para todo o $x \in [0, 1]$. Então para cada n temos

$$\left| n \int_0^1 x^{n+2} g(x) dx - n \int_0^1 x^{n+2} f(x) dx \right| = \left| n \int_0^1 x^{n+2} r(x) dx \right| \leq \varepsilon \cdot n \int_0^1 x^{n+2} dx$$

e assim, pondo $n \rightarrow \infty$ obtemos

$$\lim_{n \rightarrow \infty} \left| g(1) - n \int_0^1 x^{n+2} f(x) dx \right| \leq \lim_{n \rightarrow \infty} \varepsilon \cdot n \int_0^1 x^{n+2} dx \leq \varepsilon.$$

Como $|g(1) - f(1)| < \varepsilon$ e isto se verifica para todo o $\varepsilon > 0$, demonstrámos (1).

Key words and phrases. Teorema de Weierstrass, Argumento probabilístico.

Publicado em *Números, cirurgias e nós de gravata: 10 anos de Seminário Diagonal no IST*, J. P. Boavida, R. P. Carpentier, L. Cruz-Filipe, P. S. Gonçalves, E. Grifo, D. Henriques, A. R. Pires (editores), IST Press, 2012.

Copyright © 2012, IST Press.

O Teorema de Weierstrass é também importante em Matemática Aplicada, pois permite aproximar valores de funções contínuas por polinómios, que são mais fáceis de armazenar na memória de um computador e de calcular o valor em qualquer ponto.

Em 1912, Bernstein encontrou outra demonstração deste teorema, utilizando ferramentas probabilísticas. Esta demonstração não só é mais simples que a original de Weierstrass como tem a vantagem de permitir encontrar uma sucessão de polinómios que aproxima a função contínua. É esta demonstração que vou apresentar neste artigo, tendo como objectivo principal demonstrar de forma quase elementar este teorema fundamental em Análise e como objectivo secundário ilustrar o facto de que argumentos probabilísticos são uma ferramenta poderosa em Análise moderna, que produzem demonstrações simples e intuitivas (depois de se estar familiarizado com a linguagem das probabilidades).

Na Secção 2 introduzo as noções de probabilidades que vou utilizar, pelo que o leitor confortável com essa área pode saltar essa secção. Na Secção 3 apresento a demonstração do Teorema de Weierstrass encontrada por Bernstein. Na Secção 4 exploro as diferentes generalizações deste teorema.

2. ALGUNS CONCEITOS EM PROBABILIDADES

O elemento básico da Teoria das Probabilidades é o *espaço amostral*, chamemos-lhe Ω . Podemos pensar no espaço amostral como o conjunto de todos os possíveis estados do universo num certo tempo futuro (a definição rigorosa é um pouco técnica). Um subconjunto $A \subset \Omega$ é um *acontecimento* e tem associada uma certa *probabilidade*, $P(A)$. Por exemplo, se se atirar uma moeda ao ar pode sair cara ou coroa. Assim, em metade dos estados de Ω sai cara e na outra metade sai coroa. É claro que antes de lançar a moeda não sabemos o que vai sair, mas podemos considerar o subconjunto $A \subset \Omega$ de todos os estados futuros nos quais aquela particular moeda saiu cara, e falar da probabilidade $P(A)$ de sair cara.

Esta função P que atribui uma probabilidade a cada acontecimento tem de satisfazer certas propriedades intuitivas: $P(\Omega) = 1$; para cada acontecimento $A \subset \Omega$ temos $P(A) \geq 0$; e se A e B são acontecimentos disjuntos (também se podem chamar incompatíveis, por exemplo sair 2 no lançamento de um dado e sair 3 no mesmo lançamento) temos $P(A \cup B) = P(A) + P(B)$.¹

Uma *variável aleatória* é uma função $X : \Omega \rightarrow \mathbb{R}$, por exemplo, o número de caras em 10 lançamentos de moedas equilibradas, que a cada estado $\omega \in \Omega$ atribui o número de caras naqueles particulares 10 lançamentos que ocorreram no estado futuro ω . Um ligeiro abuso de notação muito útil é escrever $\{X = x\}$ para denotar o conjunto $\{\omega \in \Omega : X(\omega) = x\}$ e $P(X = x)$ para denotar $P(\{X = x\})$.

Todas as variáveis aleatórias utilizadas neste artigo têm um número finito de valores possíveis (o conjunto $X(\Omega)$ é finito). Assim, expressões como $\sum_{a \in \mathbb{R}} P(X = a)$ têm sentido e torna-se mais fácil desenvolver a teoria. Quando lidamos com variáveis aleatórias gerais, as definições são tecnicamente mais elaboradas, mas as ideias são as mesmas. Por exemplo, para um subconjunto $A \subset \mathbb{R}$, a seguinte soma faz sentido: $\sum_{a \in A} P(X = a) = P(X \in A)$. A igualdade verifica-se porque os acontecimentos $\{X = a\}$ são disjuntos dois a dois.

A *distribuição* de uma variável aleatória é o conjunto de valores que ela pode tomar, juntamente com as probabilidades associadas. Duas variáveis aleatórias diferentes podem ter a mesma distribuição. Por exemplo, o número X de caras no lançamento de 10 moedas e o número Y de coroas no lançamento de 10 moedas são

¹Na realidade estes são os axiomas de probabilidade finitamente aditiva, mas todos os espaços neste artigo são finitos, pelo que esta construção é equivalente.

variáveis aleatórias diferentes. No entanto, por simetria, é fácil verificar que para cada $a \in \{0, 1, \dots, 10\}$ temos $P(X = a) = P(Y = a)$.

Dois acontecimentos $A, B \subset \Omega$ dizem-se *independentes* se a informação sobre a ocorrência de um não altera a probabilidade de o outro ocorrer. Não é difícil verificar que este conceito pode ser traduzido matematicamente da seguinte maneira: os acontecimentos A e B são independentes se $P(A \cap B) = P(A)P(B)$. Por exemplo, ao tirar uma carta ao acaso de um baralho de 40 cartas, os eventos “ser paus” e “ser rei” são independentes. Da mesma forma, se X e Y são variáveis aleatórias, estas dizem-se *independentes* se qualquer informação sobre o valor de X não altera a distribuição de Y . Rigorosamente, X e Y são independentes se os acontecimentos $\{X \in A\}$ e $\{Y \in B\}$ são independentes para quaisquer $A, B \subset \mathbb{R}$. Por exemplo, ao tirar uma carta ao acaso de um baralho de 40 cartas, as variáveis “naipes” e “número” são independentes.

Uma medida muito útil associada a uma variável aleatória X é a sua *esperança* (ou valor esperado) $E[X]$, ou média. A definição deve ser intuitiva: se X é uma variável aleatória que toma apenas um número finito de valores diferentes, definimos $E[X] = \sum_{a \in \mathbb{R}} a P(X = a)$. Repare-se que as parcelas só não são zero um número finito de vezes, pelo que a soma faz sentido. Por exemplo, se X é o número de caras no lançamento de duas moedas, $E[X] = 0 \times P(X = 0) + 1 \times P(X = 1) + 2 \times P(X = 2) = 0.5 + 2 \times 0.25 = 1$. Uma propriedade fundamental da esperança é ser linear, ou seja, se X e Y são variáveis aleatórias e α é um número real, temos $E[X + Y] = E[X] + E[Y]$ e $E[\alpha X] = \alpha E[X]$.

Suponhamos que uma variável aleatória $X : \Omega \rightarrow \mathbb{R}$ é positiva e limitada, i.e., existe um M tal que $P(0 \leq X \leq M) = 1$. Então, para cada $\varepsilon \in]0, M[$ temos

$$E[X] = \sum_{a \in \mathbb{R}} a P(X = a) = \sum_{a < \varepsilon} a P(X = a) + \sum_{a \geq \varepsilon} a P(X = a).$$

Na primeira soma temos $a < \varepsilon$ e na segunda podemos assumir que $a \leq M$ (porque se $a > M$ temos $P(X = a) = 0$). Assim, obtemos a estimativa

$$(2) \quad E[X] \leq \varepsilon P(X \leq \varepsilon) + M P(X > \varepsilon).$$

É fácil ver que $|x + y| \leq |x| + |y|$, facto a que por vezes se chama *desigualdade triangular*. Podemos generalizar este facto para somas finitas e, utilizando este princípio, obtemos, para qualquer variável aleatória $X : \Omega \rightarrow \mathbb{R}$,

$$|E[X]| = \left| \sum_{a \in \mathbb{R}} a P(X = a) \right| \leq \sum_{a \in \mathbb{R}} |a P(X = a)| = \sum_{a \in \mathbb{R}} |a| P(X = a) = E[|X|].$$

Temos assim a desigualdade triangular para variáveis aleatórias: $|E[X]| \leq E[|X|]$.

Finalmente, apresentamos a clássica *desigualdade de Markov*. Não a vamos utilizar directamente, mas esta desigualdade facilita a prova da desigualdade de Chebyshev. Seja X uma variável aleatória que só toma valores não negativos. Seja Y a variável aleatória definida por $Y(\omega) = \lambda$ se $X(\omega) > \lambda$ e $Y(\omega) = 0$ caso contrário. Assim, claramente $Y(\omega) \leq X(\omega)$ para todo o $\omega \in \Omega$ e portanto $E[Y] \leq E[X]$. Por outro lado, $E[Y] = 0 \times P(Y = 0) + \lambda \times P(Y = \lambda) = \lambda P(X > \lambda)$, ou seja, $P(X > \lambda) = E[Y]/\lambda$. Finalmente, concluímos que

$$(3) \quad P(X > \lambda) \leq \frac{E[X]}{\lambda}.$$

Outra medida muito importante associada a uma variável aleatória é a sua *variância*, que nos diz quão concentrada está a variável aleatória em volta da sua

esperança. Ou seja, dá-nos uma noção de quão grande é $|X - E[X]|$. Por questões técnicas, a variância define-se utilizando o quadrado dessa variável aleatória² e assim $\text{Var}(X) = E[(X - E[X])^2]$.

A variância pode (e vai) ser usada para estimar a probabilidade de uma variável aleatória estar perto da média, através da desigualdade de Chebyshev. Seja X uma variável aleatória e seja $Y = (X - E[X])^2$. Para cada $\lambda > 0$, podemos aplicar a desigualdade de Markov (3) a Y e ficamos com $P(Y > \lambda) < E[Y]/\lambda$. Definimos $\delta = \sqrt{\lambda}$ e assim temos que os acontecimentos $\{Y > \lambda\}$ e $\{|X - E[X]| > \delta\}$ são o mesmo, logo têm a mesma probabilidade. Para além disso, $E[Y] = \text{Var}(X)$. Pondo tudo junto, temos a *desigualdade de Chebyshev*:

$$(4) \quad P(|X - E[X]| > \delta) < \frac{\text{Var}(X)}{\delta^2}.$$

Para aplicar a desigualdade de Chebyshev, é necessário calcular a variância. Sejam X e Y duas variáveis aleatórias. Ao contrário do que podíamos desejar, nem sempre $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. Contudo, se as variáveis forem independentes esta relação é válida. Usando a definição e a linearidade da esperança temos $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2E[(X - E[X])(Y - E[Y])]$. Quando X e Y são independentes temos:

$$\begin{aligned} E[XY] &= \sum_{a \in \mathbb{R}} a P(XY = a) = \sum_{x, y \in \mathbb{R}} xy P(X = x) P(Y = y) \\ &= \sum_{x \in \mathbb{R}} x P(X = x) \sum_{y \in \mathbb{R}} y P(Y = y) = E[X] E[Y]. \end{aligned}$$

Aplicando às variáveis $X - E[X]$ e $Y - E[Y]$, que são também independentes, concluímos que $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. Por outro lado, é claro pela definição que, se α é um número real e X uma variável aleatória, então $\text{Var}(\alpha X) = \alpha^2 \text{Var}(X)$.

3. DEMONSTRAÇÃO DO TEOREMA DE WEIERSTRASS

A ideia geral da demonstração é construir, para cada $x \in [0, 1]$, uma variável aleatória S_x , que tem uma grande probabilidade de estar muito perto de $f(x)$. Então a esperança $E[S_x]$ também tem de estar perto de $f(x)$. Se conseguirmos construir S_x de forma que $E[S_x]$ seja um polinómio em x , provamos o teorema. A ideia crucial é que, se $S_x = f(B_x)$ onde B_x é uma variável aleatória que tem uma grande probabilidade de estar muito perto de x , então S_x vai estar perto de $f(x)$, e se $E[B_x]$ é um polinómio em x e os possíveis valores de B_x não dependem de x , então $E[S_x]$ é também um polinómio em x . Temos no entanto de quantificar a noção de “estar muito perto”, para provarmos que o polinómio $E[S_x]$ está de facto próximo de f no sentido do teorema de Weierstrass. Repare-se que, usando esta estratégia, todo o trabalho está em construir as variáveis B_x , que são independentes de f .

Chamamos a atenção para o facto de estarmos a lidar com variáveis aleatórias que dependem de x , por exemplo, para cada parâmetro $x \in [0, 1]$ temos a variável aleatória S_x , mas quando falamos de probabilidade, esperança, etc., consideramos x fixo e S_x como uma função do espaço amostral Ω para \mathbb{R} . Nesse caso, a probabilidade, esperança, etc., será um número que depende de x (mas não de $\omega \in \Omega$).

Para cada $x \in [0, 1]$ e $k \in \mathbb{N}$ definimos a variável aleatória $b_{x,k}$ por $P(b_{x,k} = 1) = x$ e $P(b_{x,k} = 0) = 1 - x$, e assumimos que estas variáveis são todas independentes.³

²Assim, a variância é o equivalente da norma L^2 da variável aleatória. A principal vantagem em relação à norma L^1 é o produto interno, que torna L^2 num espaço de Hilbert.

³Não é imediato que existe um espaço amostral Ω e variáveis aleatórias $b_{x,k}$ com as propriedades pretendidas. No entanto, pode-se provar que conseguimos sempre encontrar variáveis aleatórias

Repare-se que $E[b_{x,k}] = 0 \times P(b_{x,n} = 0) + 1 \times P(b_{x,n} = 1) = x$. Agora, seja $B_{n,x} = \frac{1}{n} \sum_{k=1}^n b_{x,k}$. Temos $E[B_{n,x}] = \frac{1}{n} \sum_{k=1}^n E[b_{x,k}] = x$.

Podíamos agora aplicar a lei dos grandes números, que neste caso nos diz que, para n suficientemente grande, $B_{n,x}$ tem uma grande probabilidade de estar muito perto de x . No entanto, aplicando este resultado directamente temos que n depende de x , e nós precisamos de encontrar algum n que funcione para todo o $x \in [0, 1]$, ou seja, temos de provar que a convergência de $B_{x,n}$ para x é uniforme.

A ideia é aplicar a desigualdade de Chebyshev. Para isso temos de calcular a variância de $B_{x,n}$. Começamos com a variância de $b_{x,n}$:

$$\begin{aligned} \text{Var}(b_{x,n}) &= E[(b_{x,n} - E[b_{x,n}])^2] = E[b_{x,n}^2 - 2xb_{x,n} + x^2] \\ &= x^2 P(b_{x,n} = 0) + (1 - 2x + x^2) P(b_{x,n} = 1) \\ &= x^2(1 - x) + (1 - x)^2x = x(1 - x). \end{aligned}$$

Como as variáveis aleatórias $b_{x,n}$ são independentes, podemos calcular a variância de $B_{x,n}$ directamente:

$$\text{Var}(B_{x,n}) = \text{Var}\left(\frac{1}{n} \sum_{k=1}^n b_{x,k}\right) = \frac{1}{n^2} \sum_{k=1}^n \text{Var}(b_{x,n}) = \frac{x(1-x)}{n}.$$

Repare-se que, para $x \in [0, 1]$, temos $x(1-x) < 1$, portanto $\text{Var}(B_{x,n}) < 1/n$. Esta estimativa não depende de x . Utilizando a desigualdade de Chebyshev (4) temos, para cada $\delta > 0$:

$$P(|B_{x,n} - x| > \delta) \leq \frac{\text{Var}(B_{x,n})}{\delta^2} < \frac{1}{n\delta^2}.$$

Assim, escolhendo $N > 1/\delta^3$ obtemos, para todo o $n > N$ e todo o $x \in [0, 1]$:

$$(5) \quad P(|B_{x,n} - x| > \delta) < \delta.$$

Seja agora f a função que estamos a tentar aproximar. Assumimos que f é contínua em $[0, 1]$. Como $[0, 1]$ é um intervalo fechado e limitado, o teorema de Heine–Borel diz-nos que f é uniformemente contínua. Isto quer dizer que, para cada $\varepsilon > 0$, existe um $\delta > 0$ tal que, para quaisquer pontos $x, y \in [0, 1]$ a uma distância menor que δ , temos $|f(x) - f(y)| < \varepsilon$. Podemos ainda assumir que $\delta < \varepsilon$, pois claramente diminuir o δ não destrói a condição pretendida. Podemos escrever isto de outra forma: $|f(x) - f(y)| > \varepsilon \Rightarrow |x - y| > \delta$. Juntando esta observação com a equação (5) temos, para todo o n suficientemente grande,

$$P(|f(B_{x,n}) - f(x)| > \varepsilon) \leq P(|B_{x,n} - x| > \delta) < \delta < \varepsilon.$$

Seja agora n tal que esta estimativa é satisfeita e seja $B_x = B_{x,n}$ e $S_x = f(B_x)$. Temos que $E[S_x]$ é um polinómio:

$$E[S_x] = \sum_{i=0}^n f\left(\frac{i}{n}\right) P(B_{x,n} = \frac{i}{n}) = \sum_{i=0}^n f\left(\frac{i}{n}\right) \binom{n}{i} x^i (1-x)^{n-i}.$$

Agora só falta concluir que de facto $E[S_x]$ está próximo de f .

Como f é uma função contínua no intervalo fechado $[0, 1]$, é limitada. Seja M um majorante de $|f(x)|$. Assim $P(|S_x - f(x)| < 2M) = 1$ e podemos aplicar a estimativa (2) para obter

$$E[|S_x - f(x)|] \leq \varepsilon P(|S_x - f(x)| \leq \varepsilon) + 2M P(|S_x - f(x)| > \varepsilon) \leq \varepsilon(1 + 2M).$$

Assim, pela desigualdade triangular, concluímos finalmente que

$$|E[S_x] - f(x)| = |E[S_x - f(x)]| \leq E[|S_x - f(x)|] < \varepsilon(1 + 2M).$$

independentes com qualquer distribuição. É esse resultado que torna argumentos probabilísticos tão flexíveis.

A aproximação a que chegámos não foi bem ε , mas conseguimos um polinómio que está à distância $\varepsilon(1 + 2M)$ de f para todo o $\varepsilon > 0$, o que é suficiente para deduzir o teorema.

4. GENERALIZAÇÕES DO TEOREMA

Notemos que esta demonstração dá uma descrição explícita dos polinómios que aproximam f , nomeadamente os polinómios $p_n(x) = E[S_x]$. Temos

$$p_n(x) = E[S_x] = \sum_{i=0}^n \binom{n}{i} x^i (1-x)^{n-i} f\left(\frac{i}{n}\right).$$

Estes polinómios são chamados *polinómios de Bernstein* de f . Provámos que $|p_n(x) - f(x)| < \varepsilon$ para todo o $x \in [0, 1]$ se n for suficientemente grande. Para determinarmos qual o mínimo n , precisamos de saber o máximo M de $|f|$ e descobrir $\delta < \varepsilon$ tal que, se $|x - y| < \delta$, então $|f(x) - f(y)| < \varepsilon/(1 + 2M)$. Com esses dados, qualquer $n > \delta^{-3}$ satisfaz a condição desejada.

Este teorema é apenas um caso simples de um teorema mais geral. Algumas generalizações intermédias obtêm-se facilmente; por exemplo, não é difícil verificar que o teorema de Weierstrass se mantém válido para funções contínuas definidas em qualquer intervalo fechado e limitado de \mathbb{R} , $[a, b]$, através da homotetia $x \mapsto (x - a)/(b - a)$. Por outro lado, o que há de especial sobre polinómios? Em 1937, Stone propôs uma generalização do Teorema de Weierstrass na qual, em vez de polinómios, podemos aproximar funções contínuas por outras álgebras de funções.

Denotamos por $C(X)$ o conjunto de funções contínuas sobre o conjunto X . Um subconjunto $A \subset C(X)$ é uma *álgebra* se for um espaço vectorial (fechado para a soma e multiplicação por escalares) que é ainda fechado para o produto (ou seja, se $f, g \in A$, então $f \cdot g \in A$). Se alguma função constante não nula pertencer a A , todas as constantes estão em A (por multiplicação por escalares) e diz-se que A contém constantes. Se para cada par de pontos distintos $x, y \in X$ existir uma função $f \in A$ tal que $f(x) \neq f(y)$, dizemos que A separa os pontos de X . Não é difícil verificar que o conjunto dos polinómios em $[0, 1]$ é uma álgebra que contém constantes e separa os pontos de $[0, 1]$.

Na generalização do teorema de Weierstrass proposta por Stone podemos substituir os polinómios por qualquer álgebra que contenha constantes e separe pontos. Para além disso, em vez de intervalos fechados em \mathbb{R} , contempla qualquer espaço de Hausdorff compacto (por exemplo, conjuntos fechados e limitados em \mathbb{R}^d ou em \mathbb{C}^d).

Se X é um espaço de Hausdorff compacto e $A \subseteq C(X)$ uma álgebra de funções que contém constantes e separa pontos, então para toda a função $f \in C(X)$ e todo o $\varepsilon > 0$ existe uma função $g \in A$ tal que, para todo o $x \in X$, temos $|f(x) - g(x)| < \varepsilon$.

É esta generalização que é mais utilizada em Matemática Pura, com aplicações por exemplo em Análise Harmónica em grupos abelianos compactos, ao garantir que a álgebra gerada pelos caracteres é uniformemente densa no espaço das funções contínuas.

BIBLIOGRAFIA

Existem várias fontes onde podem obter mais informações sobre os assuntos discutidos neste artigo. Um boa introdução em teoria das probabilidades é [1] ou [2]. Os teoremas de cálculo utilizados podem ser encontrados em qualquer introdução a cálculo, por exemplo em [3]. Sobre o Teorema de Weierstrass e suas generalizações existe uma boa discussão em [4].

REFERÊNCIAS

- [1] Charles Grinstead e J. Laurie Snell, *Introduction to Probability*, second edition, American Mathematical Society, 1997.
- [2] Esmeralda Gonçalves e Nazaré Mendes Lopes, *Probabilidades: Princípios teóricos*, Escolar Editora, 2000.
- [3] Michael Spivak, *Calculus*, fourth edition, Publish or Perish, 2008.
- [4] Walter Rudin, *Functional Analysis*, McGraw-Hill, 1973.