Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Probabilidade em alta dimensão com aplicações em ciência de dados

Felipe Luis Giacomini

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Probabilidade em alta dimensão com aplicações em ciência de dados

Felipe Luis Giacomini
Orientador: Thiago Rodrigo Ramos

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Dezembro de 2025

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

High dimensional probability with applications in data science

Felipe Luis Giacomini
Advisor: Thiago Rodrigo Ramos

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
December 2025

Felipe Luis Giacomini

Probabilidade em alta dimensão com aplicações em ciência de dados

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por nome do(a) aluno(a) e aprovado pela banca examinadora.

Aprovado em 24 de dezembro de 2025.

Banca Examinadora:

- Prof. Dr. Thiago Rodrigo Ramos
- Prof. Dr. Luis Ernesto Bueno Salasar
- Prof. Dr. Márcio Alves Diniz

Aos meus familiares e amigos por todo suporte

Agradecimentos

Agradeço principalmente à minha família, pelo constante suporte e incentivo nos anos de escola e de faculdade. Agradeço ao meu orientador, pela enorme paciência e didática em me apresentar e ensinar tudo o que sei sobre probabilidade em alta dimensão. Agradeço aos meus professores por cada aula, ideia e recomendação que fez parte da minha formação. Por fim, agradeço aos meus amigos, pelas conversas, ajudas, risadas, almoços, trocas de ideias, tardes ociosas e o companheirismo e alegria pelos quais lembrarei dos anos de graduação.

"Part of the journey is the end."

Resumo

Com o avanço do aprendizado de máquina, novos desafios surgiram na criação e análise de algoritmos preditivos eficientes. Um dos principais obstáculos é a maldição da dimensionalidade, na qual o aumento do número de variáveis compromete o desempenho dos modelos usuais. Essa perda de poder preditivo decorre do comportamento das variáveis aleatórias em altas dimensões, fenômeno que pode ser explicado por meio de resultados da probabilidade e da teoria do aprendizado estatístico.

Neste trabalho, apresentamos e demonstramos esses fundamentos, destacando o papel central da concentração de medida em suas deduções. Para tornar a compreensão mais intuitiva, realizamos simulações que ilustram esses conceitos. Por fim, discutimos como essa teoria pode embasar métodos capazes de mitigar os desafios impostos pela alta dimensionalidade.

Palavras-chave: *concentração de medida, alta dimensão, probabilidade, aprendizado de máquinas.*

Abstract

With the advancement of machine learning, new challenges have arisen in the creation and analysis of efficient predictive algorithms. One of the main obstacles is the curse of dimensionality, in which the increase in the number of variables compromises the performance of usual models. This loss of predictive power arises from the behavior of random variables in high dimensions, a phenomenon that can be explained through results from probability and statistical learning theory.

In this work, we present and demonstrate these foundations, highlighting the central role of measure concentration in their deductions. To make understanding more intuitive, we perform simulations that illustrate these concepts. Finally, we discuss how this theory can support methods capable of mitigating the challenges imposed by high dimensionality.
Keywords: *concentration of measure, high dimension, probability, machine learning.*

Lista de Figuras

Sumário

Capítulo 1

Introdução

Com o avanço das aplicações em ciência de dados, aprendizado de máquina e estatística computacional, tornou-se cada vez mais comum a análise de dados com dezenas, centenas ou até milhares de variáveis. Essa explosão da dimensionalidade exige uma compreensão mais profunda do comportamento de objetos matemáticos em espaços de alta dimensão — em particular, vetores e matrizes aleatórios.

Diferentemente da intuição construída em baixa dimensão, muitos fenômenos geométricos e probabilísticos em $\mathbb{R}^n$ passam a exibir comportamentos qualitativamente diferentes quando n é grande. Esses efeitos — como o fato de vetores aleatórios independentes serem quase ortogonais ou de funções de vetores aleatórios variarem muito pouco — são, em grande parte, explicados pelo fenômeno de *concentração da medida* (Boucheron *et al.*, 2003). Em altas dimensões, a massa de distribuições razoavelmente regulares se concentra fortemente em torno de conjuntos de medida elevada, fazendo com que quantidades de interesse permaneçam próximas de seus valores típicos. Essas propriedades têm implicações diretas na análise e no desempenho de algoritmos de aprendizado, especialmente daqueles baseados em distâncias, ângulos ou projeções (Vershynin, 2018).

Este fenômeno também está relacionado ao que se conhece como *maldição da dimensionalidade* (Bellman, 1966; Hastie *et al.*, 2009), termo cunhado para descrever o impacto negativo que a alta dimensão pode ter sobre métodos estatísticos. Em espaços de alta dimensão, os dados tornam-se esparsos, e a concentração geométrica faz com que a maioria dos vetores estejam situados em uma fina casca ao redor de uma esfera, dificultando a diferenciação entre observações e afetando a eficácia de métodos que se baseiam em vizinhanças ou medidas de similaridade (Vershynin, 2018).

Neste trabalho, estudamos essas propriedades com mais profundidade por meio de

ferramentas de probabilidade em alta dimensão, concentrando-nos em resultados não assintóticos — isto é, válidos mesmo para valores fixos de n, embora com maior relevância à medida que n cresce. Analisamos principalmente vetores com coordenadas subgaussianas, devido às suas propriedades de concentração favoráveis, e vetores isotrópicos, que permitem uma caracterização geométrica mais simples.

A abordagem adotada neste trabalho concentra-se em três problemas fundamentais em ciência de dados, cada um explorado em um capítulo distinto.

Problema 1: Controle do erro de generalização. No Capítulo 2, estudaremos desigualdades de concentração para variáveis aleatórias. O objetivo é compreender como essas desigualdades podem ser aplicadas à análise de **aprendizado supervisionado** (Mohri *et al.*, 2018; Shalev-Shwartz e Ben-David, 2014; Devroye e Lugosi, 2001), em que se dispõe de uma classe de funções para predizer uma variável resposta a partir de um conjunto de covariáveis, minimizando uma função de perda. Nesse contexto, surge o desafio do *overfitting* (Hastie *et al.*, 2009), caracterizado por modelos que se ajustam excessivamente aos dados de treinamento, mas têm baixo poder preditivo em novas observações. Mostraremos que, quando a classe de funções é finita e o tamanho da amostra é suficientemente grande, o *overfitting* pode ser evitado com alta probabilidade.

Problema 2: Geometria de vetores em alta dimensão. No Capítulo 3, analisaremos desigualdades de concentração aplicadas a vetores aleatórios. Investigaremos dois fenômenos centrais: (i) a concentração da norma de um vetor em uma fina casca esférica, cujo raio depende da dimensão, e (ii) a quase ortogonalidade entre vetores independentes, resultante da concentração do produto interno em torno de zero (Vershynin, 2018; Wainwright, 2019a). Esses efeitos ilustram o comportamento geométrico contraintuitivo típico de espaços de alta dimensão.

Problema 3: Detecção de comunidades em redes. No Capítulo 4, introduziremos o conceito de matrizes aleatórias e apresentaremos um resultado de concentração para seus valores singulares. Esse resultado servirá de base para o estudo da detecção de comunidades em redes, no contexto do modelo de blocos estocásticos (Vershynin, 2018; Tao, 2012). Nesse modelo, os vértices são divididos em duas comunidades, e cada par de vértices é conectado, independentemente, com probabilidade p se pertencem à mesma comunidade e com probabilidade q caso contrário. Por meio de um argumento espectral,

demonstraremos o funcionamento de um algoritmo capaz de identificar corretamente as comunidades de cada vértice .

Nos capítulos seguintes, o leitor encontrará uma exposição teórica acompanhada de intuições geométricas e ilustrações obtidas por meio de simulações, de modo a reforçar o entendimento dos fenômenos estudados. Os códigos com as simulações estarão disponíveis em https://github.com/felgiacomini/tcc_prob. Espera-se que este material sirva como uma introdução acessível aos principais conceitos de probabilidade em alta dimensão e suas aplicações em ciência de dados.

Capítulo 2

Variáveis independentes e *overfitting*

Neste capítulo, motivaremos o estudo e demonstraremos algumas desigualdades de concentração para somas de variáveis aleatórias independentes. Essas desigualdades fornecem limites superiores não assintóticos — isto é, válidos para um número fixo de variáveis — para a probabilidade de a soma desviar de seu valor esperado por uma certa quantia. As desigualdades de concentração apresentadas aqui servem como introdução e motivação ao tema, além de constituírem a base para o resultado principal deste capítulo, que mostra como tais ferramentas permitem explicar problema de *overfitting*, de grande relevância na área de *machine learning* (Hastie *et al.*, 2009; Mohri *et al.*, 2018).

A Seção 2.1 traz uma breve revisão de desigualdades clássicas da teoria da probabilidade e de suas aplicações em estatística. Em seguida, na Seção 2.2, apresentamos e provamos a desigualdade de Hoeffding, ilustrando seu uso em um exemplo de algoritmos aleatórios e discutindo um refinamento específico para variáveis Bernoulli. Munidos dessa ferramenta, chegamos à Seção 2.4, em que exploramos como aplicar essas desigualdades para quantificar o tamanho amostral necessário a fim de evitar o *overfitting*.

2.1 Motivação

O estudo das desigualdades de concentração tem origem em resultados clássicos da teoria da probabilidade, que buscam quantificar a probabilidade de uma variável aleatória assumir valores distantes de sua média (Boucheron *et al.*, 2003). Essas desigualdades são fundamentais porque fornecem limites gerais para a ocorrência de eventos raros, isto é, para a probabilidade de grandes desvios em relação ao comportamento esperado.

Como ponto de partida, recordaremos algumas desigualdades básicas, amplamente

conhecidas em cursos introdutórios de probabilidade, que servirão de base para resultados mais sofisticados apresentados nas seções seguintes. A primeira delas é a *desigualdade de Markov*, que estabelece um limite superior elementar para a probabilidade de uma variável aleatória não negativa exceder um determinado valor.

Teorema 2.1 (Desigualdade de Markov)**.** *Seja X uma variável aleatória não negativa. Então, para todo $t > 0$, temos que*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Demonstração. Usando a hipótese de que $X \geq 0$, podemos limitar $\mathbb{E}[X]$ inferiormente fazendo

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X\mathbb{1}(X \geq t)] + \mathbb{E}[X\mathbb{1}(X < t)] \\
&\geq t\mathbb{E}[\mathbb{1}(X \geq t)] + 0 \\
&= t\mathbb{P}(X \geq t), && (\mathbb{E}[\mathbb{1}(X \geq t)] = \mathbb{P}(X \geq t))
\end{aligned}$$

a desigualdade é verdadeira pois $X\mathbb{1}(X \geq t) \geq t\mathbb{1}(X \geq t)$ e $\mathbb{E}[X\mathbb{1}(X < t)] \geq 0$ pois $X \geq 0$. Dividindo por t de ambos os lados retorna o resultado desejado. □

A desigualdade de Markov é válida em casos bastante gerais, assumindo apenas que $X \geq 0$ possui média finita. No entanto, ela é bastante conservadora e pode não ser muito informativa. Por outro lado, se assumirmos também que a variância de X é finita, podemos demonstrar um resultado um pouco mais forte, a desigualdade de Chebyshev, também clássica em probabilidade.

Teorema 2.2 (Desigualdade de Chebyshev)**.** *Seja X uma variável aleatória com média μ e variância $\sigma^2 < \infty$. Então, para todo $t > 0$, temos que*

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Demonstração. Seja X uma variável aleatória com média μ e variância σ^2. Então, para

todo $t > 0$,

$$\begin{aligned}\mathbb{P}(|X-\mu| \geq t) &= \mathbb{P}((X-\mu)^2 \geq t^2) \\ &\leq \frac{\mathbb{E}\left[(X-\mu)^2\right]}{t^2} && \text{(Teorema 2.1)} \\ &= \frac{\sigma^2}{t^2}.\end{aligned}$$

□

A partir da desigualdade de Chebyshev, podemos deduzir um dos resultados mais importantes da estatística: a **lei fraca dos grandes números**. De acordo com esse teorema, a média amostral $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$ converge em probabilidade para o valor esperado μ das variáveis aleatórias i.i.d. X_i. De fato, sendo σ^2 a variância comum das X_i, temos

$$\begin{aligned}\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) &\leq \frac{\operatorname{Var}\left(\bar{X}_n\right)}{t^2} && \text{(Teorema 2.2)} \\ &= \frac{\sigma^2}{nt^2} \\ &\longrightarrow 0, && \text{quando } n \to \infty.\end{aligned}$$

Esse resultado mostra que a velocidade de convergência da média amostral é, no mínimo, da ordem de $1/n$. Surge então uma questão natural: *será essa a melhor taxa possível?* Em outras palavras, é possível obter limitantes mais rígidos para a probabilidade de desvios significativos?

As desigualdades de concentração que apresentaremos nas próximas seções respondem afirmativamente a essa pergunta. Ao impor condições adicionais sobre as variáveis aleatórias em estudo — por exemplo, restrições sobre seus momentos ou suas caudas —, obteremos desigualdades muito mais fortes, com limitantes **exponencialmente decrescentes em** n para a probabilidade de a variável se desviar de sua média.

2.2 Desigualdade de Hoeffding

Nesta seção iremos provar a desigualdade de Hoeffding para variáveis limitadas. Começaremos demonstrando um lema preliminar, que será útil para limitar a função geradora de momentos das variáveis consideradas.

Lema 2.1 (Lema de Hoeffding)**.** *Seja X uma variável aleatória tal que $X \in [a, b]$ com $a, b \in \mathbb{R}$ e $\mathbb{E}[X] = 0$. Então, para todo $\lambda > 0$, temos que*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda^2 (b-a)^2/8}.$$

Demonstração. Note que $x \mapsto e^{\lambda x}$ é uma função convexa (pois $(e^{\lambda x})'' > 0$). Uma propriedade útil da convexidade é que, se $x \in [a, b]$, temos que

$$e^{\lambda x} \leq \frac{x-a}{b-a} e^{\lambda b} + \frac{b-x}{b-a} e^{\lambda a}.$$

Da expressão acima, nosso objetivo é chegar na função geradora de momentos de X, para introduzir a esperança nessa expressão, usamos o seguinte fato:

$$X \leq Y \implies \mathbb{E}[X] \leq \mathbb{E}[Y].$$

Assim, obtemos que

$$\begin{aligned} \mathbb{E}\left[e^{\lambda X}\right] &\leq \mathbb{E}\left[\frac{X-a}{b-a} e^{\lambda b} + \frac{b-X}{b-a} e^{\lambda a}\right] \\ &= \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b}. \qquad (\mathbb{E}[X] = 0) \end{aligned}$$

Fazendo a substituição de variável $t = b/(b-a)$ e $y = \lambda(b-a)$, é possível definir a função

$$\begin{aligned} \phi(y) &= \log(te^{\lambda a} + (1-t)e^{\lambda b}) \\ &= \log(e^{\lambda a}) + \log(t + (1-t)e^{\lambda(b-a)}) \\ &= \lambda(b-a)\left(\frac{b}{b-a} - 1\right) + \log(t + (1-t)e^{\lambda y}) \\ &= y(t-1) + \log(t + (1-t)e^{\lambda y}). \end{aligned}$$

Podemos limitar a função ϕ usando o teorema de Taylor, segundo o qual existe algum $c \in [0, 1]$ tal que

$$\phi(y) = \phi(0) + \phi'(0)y + \frac{\phi''(c)}{2} y^2. \tag{2.1}$$

Observe que $\phi(0) = 0$ e $\phi'(y) = t - 1 + (1-t)e^y/(t + (1-t)e^y)$ também é igual a 0 quando

$y = 0$. Para o termo da segunda derivada, temos

$$\phi''(y) = \frac{t(1-t)e^y}{t + (1-t)e^y},$$

se derivarmos novamente, descobrimos que o ponto de máximo de $\phi''(y)$ ocorre em $y_t = \log(t/(1-t))$, em que o valor máximo correspondente é $\phi''(y_t) = 1/4$. Incorporando este fato em 2.1, chegamos em

$$\phi(y) = \frac{\phi''(c)}{2} y^2 \leq \frac{y^2}{8} = \frac{\lambda^2 (b-a)^2}{8},$$

como queríamos demonstrar. □

Munidos do Lema 2.1, estamos prontos para enunciar a **desigualdade de Hoeffding**, principal resultado desta seção. Sua demonstração emprega uma técnica amplamente utilizadas em concentração de medida: o *método de Cramér–Chernoff*. Essa abordagem baseia-se em limitar probabilidades por meio da função geradora de momentos, fornecendo um caminho sistemático para derivar desigualdades exponenciais.

Teorema 2.3 (Desigualdade de Hoeffding)**.** *Sejam $X_1, \ldots, X_n$ variáveis aleatórias independentes em que $X_i \in [a_i, b_i]$ para todo i. Então, para todo $t > 0$, temos que*

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \right| \geq t \right) \leq 2 \exp\left(\frac{-2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

Demonstração. Começaremos a demonstração considerando apenas a cauda superior; a generalização para a cauda inferior será análoga. Para isso, note que, para todo $\lambda > 0$,

temos

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right) &= \mathbb{P}\left(\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right) \geq e^{\lambda t}\right) \\
&\leq e^{-\lambda t}\,\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right)\right] && \text{(Teorema 2.1)} \\
&= e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] && \text{(independência)} \\
&\leq e^{-\lambda t}\prod_{i=1}^{n}e^{\lambda^2(b_i-a_i)^2/8} && \text{(Lema 2.1)} \\
&= \exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2 - \lambda t\right).
\end{aligned}$$

Na primeira igualdade acima, multiplicamos ambos os lados da desigualdade dentro da probabilidade por λ e aplicamos a função exponencial. Como $y \to e^{\lambda y}$ é uma função estritamente crescente para todo $\lambda > 0$, então $e^{y_1} \geq e^{y_2}$ se e somente se $y_1 \geq y_2$. Essa transformação é o passo inicial do *método de Cramér–Chernoff*, que permite converter uma probabilidade em uma expectativa exponencial, facilitando o uso de desigualdades como a de Markov.

Uma vez que o valor de λ é arbitrário, podemos minimizar em λ para obter o melhor limitante possível. No caso, este ponto de mínimo ocorre para $\lambda_t = 4t/\sum_{i=1}^{n}(b_i - a_i)^2$ e resulta em

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}\left[X_i\right]) \geq t\right) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Por fim, para a provar a desigualdade considerando o valor absoluto, observe que para qualquer variável aleatória Y e para todo $t > 0$, temos que

$$\mathbb{P}(|Y| \geq t) = \mathbb{P}(Y \geq t) + \mathbb{P}(Y \leq -t).$$

Portanto, basta tomarmos $Y = \sum_{i=1}^{n}(X_i - \mathbb{E}\left[X_i\right])$ na equação acima e aplicarmos a desigualdade já demonstrada para $\mathbb{P}(Y \leq -t) = \mathbb{P}(-Y \geq t)$, o que é válido pois $-Y$ também será uma variável limitada, isto completa a prova. □

Exemplo 2.1. *Como aplicação da desigualdade de Hoeffding, consideremos um exemplo em uma área fundamental da ciência da computação: os* ***algoritmos aleatorizados*** *(Mohri et al., 2018). Suponha que desejamos resolver um problema de decisão (por exem-*

plo, determinar se um número p é primo). Considere um algoritmo que, a cada execução, faz uma escolha aleatória e fornece a resposta correta com probabilidade $1/2 + \delta$*, para algum* $\delta > 0$*.*

Para aumentar a confiabilidade da resposta, podemos executar o algoritmo n vezes de forma independente e adotar a decisão por maioria*. Intuitivamente, quanto maior o número de repetições, menor será a probabilidade de a maioria das respostas estar incorreta. Mostraremos que, para qualquer* $\varepsilon \in (0, 1)$*, a decisão será correta com probabilidade pelo menos* $1 - \varepsilon$*, desde que n seja suficientemente grande em função de* δ *e* ε*:*

$$n \geq \frac{1}{2\delta^2} \log\left(\frac{1}{\varepsilon}\right).$$

Com este objetivo, defina as variáveis aleatórias $\{X_i\}_{i=1}^n$ *independentes e identicamente distribuídas tais que, para todo i:*

$$X_i = \begin{cases} 1, & \text{se o algoritmo retornar a resposta errada na } i\text{-ésima iteração} \\ 0, & \text{caso contrário.} \end{cases}$$

Isto é, as indicadoras de resposta errada em cada uma das n vezes em que rodaremos o algoritmo. A resolução consiste em garantir que a probabilidade de escolher a resposta errada após tomar o voto da maioria seja no máximo ε*, isto é,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq \frac{n}{2}\right) \leq \varepsilon, \tag{2.2}$$

para isso, aplicamos a desigualdade de Hoeffding. Note que

$$\mathbb{E}\left[X_i\right] = 0 \cdot (1/2 + \delta) + 1 \cdot (1 - (1/2 + \delta)) = 1/2 - \delta,$$

para todo $1 \leq i \leq n$*. Portanto, subtraindo* $n(1/2 - \delta)$ *dos dois lados da probabilidade em*

2.2, obtemos:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq \frac{n}{2}\right) &= \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}\left[X_i\right]) \geq \frac{n}{2} - \frac{n(1-2\delta)}{2}\right) \\
&= \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}\left[X_i\right]) \geq \delta n\right) \\
&\leq \exp\left(\frac{-2(\delta n)^2}{\sum_{i=1}^{n}(1-0)^2}\right) \qquad \text{(Teo 2.3)} \\
&= \exp\left(-2\delta^2 n\right).
\end{aligned}$$

Queremos que o limitante acima seja menor ou igual a ε, para isto, é fácil ver que basta tomar $n \geq (-1/2\delta^2)\log(\varepsilon)$, como queríamos mostrar.

2.3 Desigualdade de Chernoff

Em alguns contextos, a desigualdade de Hoeffding pode se mostrar excessivamente conservadora, produzindo limites pouco informativos. Um exemplo clássico ocorre quando consideramos uma sequência de variáveis aleatórias independentes $X_1, X_2, \ldots, X_n$, em que cada X_i segue uma distribuição de Bernoulli com parâmetro p_i.

Suponha que os parâmetros p_i decaem para zero de forma suficientemente rápida, de modo que exista $\lambda > 0$ tal que

$$\sum_{i=1}^{n} p_i \longrightarrow \lambda, \quad \text{quando } n \to \infty.$$

Nesse regime, pode-se demonstrar que a soma $S_n = \sum_{i=1}^{n} X_i$ converge em distribuição para uma variável aleatória Poisson com parâmetro λ. Assim, a verdadeira cauda da distribuição de S_n é muito mais leve do que aquela sugerida pela desigualdade de Hoeffding, que ignora o fato de as variáveis assumirem apenas valores binários.

O próximo resultado, conhecido como **desigualdade de Chernoff**, fornece um limitante exponencial mais ajustado para o caso de somas de variáveis de Bernoulli independentes, refletindo de maneira mais fiel o comportamento de suas caudas.

Teorema 2.4. *Sejam $\{X_i\}_{i=1}^{n}$ variáveis aleatórias independentes com distribuição de Ber-*

noulli de parâmetros p_i e considere $\mu = \sum_{i=1}^{n} p_i$. Então, para todo $t > 0$, temos que

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq e^{t-\mu}\left(\frac{\mu}{t}\right)^t.$$

Demonstração. Iremos novamente usar o método de Crámer-Chernoff para limitar a probabilidade desejada, assim, para todo $\lambda > 0$, pela Desigualdade de Markov 2.1,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda X_i}\right],$$

em que podemos limitar as esperanças no produtório usando a desigualdade útil do Apêndice A.3 da seguinte forma:

$$\mathbb{E}\left[e^{\lambda X_i}\right] = p_i e^{\lambda} + 1 - p_i = 1 + (e^{\lambda} - 1)p_i \leq e^{(e^{\lambda}-1)p_i}.$$

Com isso, concluímos que o limitante será

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) &\leq e^{-\lambda t}\prod_{i=1}^{n}\exp((e^{\lambda}-1)p_i)\\ &\leq \exp\left((e^{\lambda}-1)\sum_{i=1}^{n} p_i - \lambda t\right).\end{aligned}$$

Minimizando com $\lambda_t = \log(t/\sum_{i=1}^{n} p_i) = \log(t/\mu)$ obtemos que

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq e^{(t/\mu-1)\mu}\left(\frac{t}{\mu}\right)^{-t} = e^{t-\mu}\left(\frac{\mu}{t}\right)^t.$$

□

2.4 Aplicação: Aprendizado estatístico e *overfitting*

As desigualdades de concentração constituem um campo de estudo amplo dentro da probabilidade, com aplicações diretas em diversas áreas, em especial na de *machine learning* (Mohri *et al.*, 2018; Shalev-Shwartz e Ben-David, 2014; Devroye e Lugosi, 2001; Vershynin, 2018; Bach, 2024; Wainwright, 2019a). Essa área da inteligência artificial tem como objetivo desenvolver algoritmos capazes de aprender a partir de informações amostrais, sem a necessidade de serem explicitamente programados para cada tarefa. Tais

algoritmos produzem modelos preditivos que buscam generalizar o comportamento observado em exemplos passados.

Um dos paradigmas centrais é o **aprendizado supervisionado**, no qual o algoritmo recebe pares de entrada e saída (exemplos rotulados) e deve aprender uma função que relacione essas variáveis (Hastie *et al.*, 2009). O objetivo é encontrar uma função que, ao receber uma nova entrada, seja capaz de prever corretamente a saída correspondente.

Seja $\mathcal{X}$ o espaço de instâncias (por exemplo, imagens ou vetores de atributos) e $\mathcal{Y}$ o espaço de rótulos discreto (por exemplo, $\{0, 1\}$ em um problema de classificação binária). O aprendizado ocorre a partir de um conjunto de treino com m pares $S = \{(X_1, Y_1), \ldots, (X_m, Y_m)\}$, em que cada par (X_i, Y_i) é amostrado i.i.d. de uma distribuição desconhecida $\mathcal{D}$ sobre $\mathcal{X} \times \mathcal{Y}$.

Definição 2.1 (Função hipótese)**.** *Denomina-se* ***hipótese*** *uma função $h : \mathcal{X} \to \mathcal{Y}$, escolhida a partir de uma* ***classe de hipóteses*** *$\mathcal{H}$. Cada função $h \in \mathcal{H}$ representa um possível modelo preditivo que o algoritmo pode selecionar com base nos dados observados.*

O conjunto $\mathcal{H}$, portanto, define o espaço de busca do algoritmo de aprendizado — isto é, o conjunto de todas as funções candidatas consideradas pelo método. O objetivo é identificar, dentro dessa classe, a hipótese que minimize o **erro de generalização**, o qual mede o desempenho de h em novas amostras provenientes da distribuição desconhecida $\mathcal{D}$.

Definição 2.2 (Erro de Generalização)**.** *O erro de generalização de uma hipótese h é dado por:*

$$L_{\mathcal{D}}(h) = \mathbb{P}_{\mathcal{D}}(h(X) \neq Y).$$

No entanto, não conhecemos a distribuição $\mathcal{D}$ que gerou os dados e o melhor que podemos fazer é estimá-la, fazemos isso através do erro empírico, que vale para cada $h \in \mathcal{H}$ possível.

Definição 2.3 (Erro Empírico)**.** *O* ***erro empírico*** *de h no conjunto de treino S é:*

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[h(x_i) \neq y_i],$$

onde $\mathbb{1}[\cdot]$ é a função indicadora.

Definição 2.4 (Algoritmo de Aprendizado)**.** *Um* ***algoritmo de aprendizado*** *é um procedimento que, dado um conjunto de treino S, retorna uma hipótese $h_S \in \mathcal{H}$.*

O objetivo central da teoria do aprendizado é compreender sob quais condições, e com quais garantias, um algoritmo pode retornar uma hipótese com baixo erro de generalização. A chave para isso está em analisar a relação entre o erro empírico e o erro verdadeiro, bem como o impacto da escolha da classe $\mathcal{H}$ sobre a capacidade de generalização do modelo.

Um dos principais desafios nesse contexto é o fenômeno de *overfitting* (Hastie *et al.*, 2009), que ocorre quando o modelo se ajusta muito bem aos dados de treinamento, mas apresenta baixo desempenho em novas observações. Em outras palavras, ao minimizarmos apenas o erro empírico, podemos obter uma função que memoriza os dados, sem capturar o padrão subjacente da distribuição.

Traduzindo esse conceito para a linguagem apresentada acima, evitar o *overfitting* equivale a garantir que o erro empírico e o erro verdadeiro sejam próximos com alta probabilidade (Mohri *et al.*, 2018; Shalev-Shwartz e Ben-David, 2014). Formalmente, desejamos que, para todo $\varepsilon > 0$, existe um $\delta > 0$ tal que, para todo $h \in \mathcal{H}$,

$$\mathbb{P}\left(|L_{\mathcal{D}}(h) - L_S(h)| \geq \varepsilon\right) \leq \delta. \tag{2.3}$$

Neste capítulo, mostraremos que, quando o conjunto de hipóteses $\mathcal{H}$ é finito, podemos garantir a desigualdade (2.3) sempre que

$$m \geq \frac{1}{2\varepsilon^2}\left(\log|\mathcal{H}| + \log\left(\frac{2}{\delta}\right)\right). \tag{2.4}$$

Para demonstrar esse resultado, utilizaremos a desigualdade de Hoeffding (Teorema 2.3). Observe que, em nosso caso, as variáveis $|L_S(h) - L_{\mathcal{D}}(h)|$ são limitadas no intervalo $[0, 1]$ e satisfazem $\mathbb{E}[L_S(h)] = L_{\mathcal{D}}(h)$. Assim, para cada $h \in \mathcal{H}$, temos

$$\mathbb{P}\left(|L_S(h) - L_{\mathcal{D}}(h)| \geq \varepsilon\right) \leq 2\exp\left(-2\varepsilon^2 m\right).$$

A partir desse resultado, podemos limitar a probabilidade de que exista algum $h \in \mathcal{H}$ que viole a condição (2.3) aplicando o **limitante da união**:

$$\begin{aligned}\mathbb{P}\left(\exists\, h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| \geq \varepsilon\right) &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(|L_S(h) - L_{\mathcal{D}}(h)| \geq \varepsilon\right) \\ &\leq |\mathcal{H}| \cdot 2\exp\left(-2\varepsilon^2 m\right).\end{aligned}$$

Substituindo (2.4) na desigualdade acima, obtemos precisamente (2.3). Assim, quando a classe de hipóteses $\mathcal{H}$ é finita, o erro de generalização é limitado, temos a garantia de que, para uma amostra suficientemente grande, com alta probabilidade, o *overfitting* não ocorre. É importante observar que, para esse resultado, nenhuma suposição adicional foi feita sobre a distribuição $\mathcal{D}$.

Para ilustrar o efeito do tamanho do conjunto de hipóteses para predição da variável, realizaremos uma simulação de exemplo, usando um conjunto de dados de classificação. A base de dados simulada é feito usando a biblioteca `sklearn` na linguagem Python. Gera-se uma base de dados de 10000 linhas com 3 covariáveis e uma variável resposta binária. Em seguida, dividimos a base em 100 blocos de 1000 linhas (700 para treino e 300 para teste). Por fim, treinamos um modelo de árvore rasa (profundidade máxima igual a 3) e um modelo de árvore profunda (sem limite de profundidade máxima) em cada bloco, registrando a acurácia no treino e acurácia no teste para cada modeloe calculando a diferença absoluta entre as duas.

A distribuição das diferenças absolutas pode ser observada na Figura 2.1. Nota-se que a diferença entre o erro no treino e o erro no teste se concentra em valores consideravelmente maiores para a árvore profunda. Isso ocorre porque a árvore profunda é um modelo que seleciona entre uma quantidade muito maior de quebras no espaço das covariáveis, isto é, possui um $|\mathcal{H}|$ maior. Por outro lado, a árvore rasa considera uma quantidade reduzida de hipóteses possíveis, ou seja, tem $|\mathcal{H}|$ menor. Assim, em acordo com o resultado que demonstramos, o *overfitting* é mais provável em um modelo que considera um conjunto de hipóteses maior, fixado o mesmo tamanho de amostra.

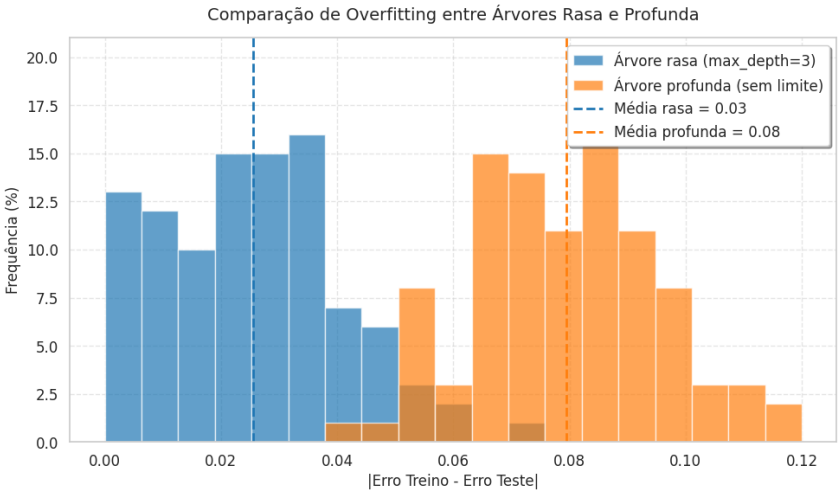

Figura 2.1: Histograma das diferenças entre erros para uma árvore rasa e uma árvore profunda.

Capítulo 3

Vetores aleatórios e maldição da dimensionalidade

Neste capítulo iremos estudar o comportamento de vetores aleatórios em $\mathbb{R}^n$. Seguindo a linha do capítulo anterior, os teoremas apresentados são não assintóticos, sendo válidos para uma dimensão fixa $n \in \mathbb{N}$, mas se tornam interessantes quando o número de dimensões é grande. No espaço euclidiano n-dimensional, temos exponencialmente mais espaço conforme n cresce, o que pode prejudicar a performance de algoritmos de predição, já que a maior esparsidade das observações dificulta o reconhecimento de padrões, um problema conhecido como maldição da dimensionalidade (Bellman, 1966; Hastie *et al.*, 2009).

Na Seção 3.1 introduziremos o importante conceito de variáveis aleatórias subgaussianas, das quais as variáveis Bernoulli estudadas no capítulo anterior são um caso especial. Na Seção 3.2, apresentamos a família de variáveis subexponenciais, das quais variáveis com caudas exponenciais como a Poisson e a Qui-Quadrado são exemplos. O espaço das variáveis subgaussianas e subexponencias é extremamente rico para o estudo de fenômenos probabilísticos em alta dimensão.

A Seção 3.3 apresentará a Desigualdade de Bernstein, que servirá de apoio para o estudo das normas de vetores aleatórios na Seção 3.4. Mostaremos também como se comportam os ângulos entre vetores aleatórios em alta dimensão na Seção 3.5. Por fim, motivados pelos resultados das seções anteriores, apresentamos e provamos o Lema de Johnson-Lindestrauss, um resultado que demonstra uma método simples de redução de dimensionalidade capaz de suavizar os efeitos apresentados nas seções anteriores.

3.1 Variáveis subgaussianas

Nesta seção iremos caracterizar as distribuições subgaussianas e provar resultados de concentração para esta classe de variáveis. Em geral, são reconhecidas por terem caudas que decaem como a cauda de uma normal, isto é, se X é uma variável subgaussiana, então, para todo $t > 0$,

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-2t^2).$$

As variáveis Bernoulli, assim como as limitadas em geral e a normal são exemplos de distribuições subgaussianas. Assim, as subgaussianas são uma generalização das variáveis aleatórias estudadas até aqui.

Antes de apresentar a definição formal que usaremos de variável subgaussiana, começaremos a seção com uma proposição que dá a equivalência entre a cauda, os momentos e a função geradora de momentos de uma variável desta classe. Com isso, será possível identificar uma variável subgaussiana a partir de qualquer uma dessas características.

Proposição 3.1. *Para uma variável aleatória X e constantes $K_1, \ldots, K_5$ que diferem entre si por no máximo um fator constante, as seguintes afirmações são equivalentes.*

1. *Existe $K_1 > 0$ tal que*

$$\mathbb{P}(|X| \geq t) \leq 2\exp(-t^2/K_1^2), \text{ para todo } t \geq 0.$$

2. *Existe $K_2 > 0$ tal que os momentos de X satisfazem*

$$(\mathbb{E}[|X|^p])^{1/p} \leq K_2\sqrt{p}, \text{ para todo } p \geq 1.$$

3. *Existe $K_3 > 0$ tal que*

$$\mathbb{E}\left[\exp(\lambda^2 X^2)\right] \leq \exp(K_3^2\lambda^2), \text{para todo } \lambda \text{ tal que } |\lambda| \leq 1/K_3.$$

4. *Existe $K_4 > 0$ tal que*

$$\mathbb{E}\left[\exp(X^2/K_4^2)\right] \leq 2.$$

5. *Se $\mathbb{E}[X] = 0$, existe $K_5 > 0$ tal que*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(K_5^2\lambda^2) \text{ para todo } \lambda \in \mathbb{R}.$$

Demonstração. (1) $\implies$ (2). Assuma que $K_1 = 1$, então

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty \mathbb{P}\left(|X| > t\right) pt^{p-1} dt && \text{(Apêndice A.2)} \\
&\leq p \int_0^\infty 2e^{-t^2} t^{p-1} dt && \text{(Afirmação (1))} \\
&= p \int_0^\infty e^{-s} s^{p/2-1} ds && (s = t^2) \\
&= p\Gamma(p/2) \\
&\leq 3p(p/2)^{p/2} \quad , && (\Gamma(x) \leq 3x^x \text{ se } x \geq 1/2)
\end{aligned}$$

elevando a $1/p$ dos dois lados chegamos ao resultado com $K_2 \leq 3$.

(2) $\implies$ (3). Assumindo $K_2 = 1$, pela expansão em série Taylor temos

$$\mathbb{E}\left[e^{\lambda^2 X^2}\right] = \mathbb{E}\left[1 + \sum_{p=1}^{\infty} \frac{(\lambda^2 X^2)^p}{p!}\right] = 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p} \mathbb{E}\left[X^{2p}\right]}{p!}.$$

Para limitar essa expressão, usamos a suposição (2), segundo a qual

$$(\mathbb{E}\left[X^{2p}\right])^{1/2p} \leq \sqrt{2p} \implies \mathbb{E}\left[X^{2p}\right] \leq 2p^p,$$

além disso, pela fórmula de Stirling, temos que $p! \geq (p/e)^p$. Portanto, temos que

$$\begin{aligned}
1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p} \mathbb{E}\left[X^{2p}\right]}{p!} &\leq 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p} (2p)^p}{(p/e)^p} \\
&= 1 + \sum_{p=1}^{\infty} (2e\lambda^2)^p \\
&= \frac{1}{1 - 2e\lambda^2} && \text{(Supondo } 2e\lambda^2 < 1\text{)} \\
&\leq e^{4e\lambda^2} \quad , && \text{(Usando A.4)}
\end{aligned}$$

em que na última desigualdade supomos adicionalmente que $2e\lambda^2 < 1/2$, o que retorna a propriedade (3) com $K_3 = 2\sqrt{e}$.

(3) $\implies$ (4). Tome $K_3 = 1$ e K_4 uma constante tal que $1/K_4^2 \leq 1$, pela propriedade (3) temos então que

$$\mathbb{E}\left[\exp(X^2/K_4^2)\right] \leq e^{(1/K_4^2)^2}.$$

Portanto adicionando a restrição $(1/K_4^2)^2 \leq \log 2$ obtemos a afirmação (4).

(4) $\implies$ (1). Assumindo $K_4 = 1$ aplicando a desigualdade de Markov:

$$\begin{aligned}
\mathbb{P}\left(|X| > t\right) &= \mathbb{P}\left(e^{X^2} \geq e^{t^2}\right) \\
&\leq e^{-t^2}\mathbb{E}\left[e^{X^2}\right] \\
&\leq 2e^{-t^2}, && \text{(Proposição (4))}
\end{aligned}$$

ou seja, a propriedade (1) com $K_1 = 1$ Para provar as equivalências da afirmação (5), consideramos $\mathbb{E}\left[X\right] = 0$.

(3) $\implies$ (5). Assuma $K_3 = 1$ e considere primeiro $|\lambda| \leq 1/K_3 = 1$. Então

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X}\right] &\leq \mathbb{E}\left[\lambda X + e^{\lambda^2 X^2}\right] && \text{(A.5)} \\
&= \mathbb{E}\left[e^{\lambda^2 X^2}\right] && (\mathbb{E}\left[X\right] = 0) \\
&\leq e^{\lambda^2}, && \text{(Proposição (3))}
\end{aligned}$$

essa é a proposição (5) com $|\lambda| \leq 1$ e $K_5 = 1$. Para $|\lambda| > 1$, temos

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda X}\right] &\leq e^{\lambda^2/2}\mathbb{E}\left[e^{X^2/2}\right] && (2\lambda X \leq \lambda^2 + X^2) \\
&\leq e^{\lambda^2/2}e^{1/2} && \text{(Proposição (3))} \\
&\leq e^{\lambda^2}, \qquad . && \text{(pois } (\lambda^2+1)/2 \leq \lambda^2)
\end{aligned}$$

em que na última desigualdade usamos a hipótese $|\lambda| > 1$. Novamente, temos a afirmação (5) com $K_5 = 1$.

(5) $\implies$ (1). Assuma $K_5 = 1$. Novamente, pela desigualdade de Markov temos, para todo $t > 0$

$$\begin{aligned}
\mathbb{P}\left(X \geq t\right) &\leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda X}\right] \\
&\leq e^{\lambda^2}e^{-\lambda t}. && \text{(Proposição (5))}
\end{aligned}$$

Minimizando com $\lambda_t = t/2$ e aplicando o mesmo processo para $-X$, obtemos

$$\mathbb{P}\left(|X| > t\right) \leq 2\exp(-t^2/4),$$

ou seja, a proposição (1) com $K_1 = 2$. □

Podemos agora definir formalmente variáveis subgaussianas, sabendo que elas irão satisfazer todas as proposições acima. Além disso, podemos associar uma norma ao espaço das variáveis subgaussianas, definida abaixo.

Definição 3.1 (Norma subgaussiana)**.** *Uma variável aleatória X é denominada subgaussiana se satisfaz qualquer uma das afirmações equivalentes da Proposição 3.1. A norma subgaussiana de X, denotada por $||X||_{\psi_2}$, é definida por*

$$||X||_{\psi_2} = \inf\{t > 0 : \mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2\}.$$

Uma propriedade útil dessas variáveis é que a soma de subgaussianas centradas é ainda subgaussiana e, além disso, a norma da soma é limitada pela soma das normas multiplicada por um fator constante. Provaremos este fato a seguir e o usaremos para deduzir uma desigualdade do mesmo tipo de Hoeffding para subgaussianas em geral.

Teorema 3.1 (Soma de subgaussianas)**.** *Sejam $X_1, \ldots, X_n$ variáveis subgaussianas independentes com média 0. Então, para algum $C > 0$ vale que*

$$\left|\left|\sum_{i=1}^{n} X_i\right|\right|_{\psi_2}^2 \leq C \sum_{i=1}^{n} ||X_i||_{\psi_2}^2.$$

Demonstração. Uma vez que $\mathbb{E}[X_i] = 0$ para todo i, iremos limitar a função geradora de momentos da soma das variáveis para chegar na afirmação (5) da Proposição 3.1. Assim, para todo $\lambda \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{n} X_i\right)\right] &= \prod_{i=1}^{n} \mathbb{E}\left[\exp(\lambda X_i)\right] \\
&\leq \prod_{i=1}^{n} \exp(C\lambda^2 ||X_i||_{\psi_2}) && \text{(Item (5) de 3.1)} \\
&= \exp(\lambda^2 K^2). && (K^2 = C\textstyle\sum_{i=1}^{n} ||X_i||_{\psi_2}^2)
\end{aligned}$$

Assim obtemos que a soma também satisfaz a afirmação (5) com norma ao quadrado menor ou igual a K^2, como queríamos demonstrar. □

Teorema 3.2. *Sejam $X_1, \ldots, X_n$ variáveis subgaussianas independentes com média 0.*

Então, existe $c > 0$ tal que para todo $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq \exp\left(\frac{-ct^2}{\sum_{i=1}^{n} ||X_i||_{\psi_2}^2}\right).$$

O seguinte resultado nos dá uma forma mais conveniente de expressar o teorema anterior.

Demonstração. De acordo com o Teorema 3.1, a soma de subgaussianas satisfaz a propriedade (2) da Proposição 3.1. Portanto, ela também satisfaz a propriedade 1, que é exatamente o que está sendo afirmado. □

Os últimos teoremas apresentados dependem das variáveis envolvidas terem média 0. Caso este não seja o caso, podemos subtrair a esperança, como mostra o resultado abaixo.

Teorema 3.3 (Centramento)**.** *Seja X uma variável aleatória subgaussiana, então $X - \mathbb{E}[X]$ também é subgaussiana e, para algum $C > 0$,*

$$||X - \mathbb{E}[X]||_{\psi_2} \leq C||X||_{\psi_2}.$$

Demonstração. Usando a propriedade da desigualdade triangular para a norma, temos que $||X - \mathbb{E}[X]||_{\psi_2} \leq ||X||_{\psi_2} + ||\mathbb{E}[X]||_{\psi_2}$. Seja $a = \mathbb{E}[X]$ e note que, como a é constante, temos $||a||_{\psi_2} = |a|/\sqrt{\log 2}$. Portanto, para sendo $c = \sqrt{\log(2)}$

$$\begin{aligned} ||\mathbb{E}[X]||_{\psi_2} &\leq c|\mathbb{E}[X]| \\ &\leq \mathbb{E}[|X|] && \text{(Desigualdade de Jensen)} \\ &\leq C||X||_{\psi_2}, && \text{((2) de 3.1)} \end{aligned}$$

este é o resultado desejado. □

3.2 Variáveis subexponenciais

Na seção anterior provamos resultados de concentração para variáveis que possuem caudas subgaussianas, no entanto, essa classe pode ser um tanto restritiva. Um exemplo disso é quando consideramos X^2 no caso em que $X \sim \mathcal{N}(0, 1)$. Essa variável terá distribuição qui-quadrado com 1 grau de liberdade e não será subgaussiana. No entanto,

ela ainda possui uma cauda que decai como a de uma distribuição exponencial. Variáveis com esta característica são chamadas de subexponenciais. A importância do estudo destas variáveis ficará mais claro quando falarmos sobre a norma euclidiana de um vetor aleatório de coordenadas subgaussianas, em que, mesmo as variáveis estando ao quadrado, ainda é possível mostrar que elas se concentram.

Estudaremos agora então esta classe de distribuições que abrange variáveis com caudas subexponenciais e provaremos um resultado a análogo para o da seção anterior sobre suas propriedades.

Proposição 3.2. *Seja X uma variável aleatória e constantes $K_1, \dots, K_5$ que diferem entre si por no máximo um fator constante, as seguintes afirmações são equivalentes.*

1. *Existe $K_1 > 0$ tal que, para todo $t > 0$*

$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp(-t/K_1).$$

2. *Existe $K_2 > 0$ tal que, para todo $p \geq 1$*

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq K_2 p.$$

3. *Existe $K_3 > 0$ tal que*

$$\mathbb{E}\left[e^{\lambda|X|}\right] \leq \exp(K_3\lambda), \text{ para } 0 \leq \lambda \leq 1/K_3.$$

4. *Existe $K_4 > 0$ tal que*

$$\mathbb{E}\left[\exp(|X|/K_4)\right] \leq 2.$$

5. *Se $\mathbb{E}\left[X\right] = 0$, existe $K_5 > 0$ tal que*

$$\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp(K_5^2\lambda^2), \text{ para } |\lambda| \leq 1/K_5.$$

Demonstração. Começaremos provando a equivalência (2) $\iff$ (5), todas as outras serão análogas ao caso subgaussiano.

(2) $\implies$ (5). Assuma $K_2 = 1$ e note que da proposição (2) e da fórmula de Stirling, temos $\mathbb{E}\left[X^p\right] \leq p^p$ e $p! \geq (p/e)^p$. Tome λ tal que $|e\lambda| \leq 1/2$. Então, pela expansão de

Taylor temos

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda X)\right] &= 1 + \sum_{p=2}^{\infty} \frac{\lambda^p \mathbb{E}\left[X^p\right]}{p!} \\ &\leq 1 + \sum_{p=2}^{\infty} \frac{\lambda^p p^p}{(p/e)^p} \\ &= 1 + \sum_{p=2}^{\infty} (e\lambda)^p \\ &= 1 + \frac{(e\lambda)^2}{1 - e\lambda}. \qquad (|e\lambda| < 1)\end{aligned}$$

Usando a suposição $|e\lambda| \leq 1/2$ e A.3 temos

$$1 + \frac{(e\lambda)^2}{1 - e\lambda} \leq 1 + 2(e\lambda)^2 \leq \exp(2e^2\lambda^2),$$

ou seja, a proposição (5) com $K_5 = 2e$.

(5) $\implies$ (2). Assuma $K_5 = 1$, pela desigualdade no Apêndice A.6, sabemos que

$$\mathbb{E}\left[|X|^p\right] \leq p^p(\mathbb{E}\left[e^X\right] + \mathbb{E}\left[e^{-X}\right]).$$

Pela propriedade (5), vale que $\mathbb{E}\left[e^X\right] \leq e$ e $\mathbb{E}\left[e^{-X}\right] \leq e$. Portanto,

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq (2e)^{1/p} p,$$

isto é, a propriedade (2) com $K_2 \leq 2e$.

(1) $\implies$ (2). Assuma $K_1 = 1$, então

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty \mathbb{P}\left(|X| > t\right) p t^{p-1} dt \leq 2p \underbrace{\int_0^\infty t^{p-1} e^{-t} dt}_{\Gamma(p)} \leq 6p^{p+1},$$

em que a última desigualdade decorre do fato de que $\Gamma(p) \leq 3p^p$ para $p \geq 1/2$. Esta é a proposição (2) com $K_2 \leq 6$.

(2) $\implies$ (3). Assuma $K_2 = 1$. Pela expansão de Taylor, temos

$$\mathbb{E}\left[e^{\lambda|X|}\right] = 1 + \sum_{p=1}^{\infty} \frac{\lambda^p \mathbb{E}\left[|X|^p\right]}{p!} \leq 1 + \sum_{p=1}^{\infty} \frac{\lambda^p p^p}{p^p/e^p} = \frac{1}{1 - e\lambda},$$

na primeira desigualdade usamos a propriedade (2) e na última igualdade assumimos $|e\lambda| < 1$. Para limitar a última expressão tomamos $0 \leq e\lambda \leq 1/2$ para obter, a partir de A.4,

$$\frac{1}{1 - e\lambda} \leq e^{2e\lambda},$$

esta é a proposição (3) com $K_3 = 2e$.

(3) $\implies$ (4). Assuma $K_3 = 1$ e tome K_4 tal que $1/K_4 \leq 1$. Então, pela propriedade (3),

$$\mathbb{E}\left[\exp(|X|/K_4)\right] \leq \exp(1/K_4).$$

Portanto, se tomarmos K_4 tal que $1/K_4 \leq \log 2$ obtemos a proposição (3).

(4) $\implies$ (1). Assuma $K_4 = 1$, então, pela desigualdade de Markov,

$$\mathbb{P}\left(|X| \geq t\right) = \mathbb{P}\left(e^{|X|} \geq e^t\right) \leq e^{-t}\mathbb{E}\left[e^{|X|}\right] \leq 2e^{-t},$$

em que na última desigualdade usamos a propriedade (4) e obtivemos a propriedade (1) com $K_1 = 1$. □

De maneira semelhante ao que foi feito na seção anterior, definiremos uma norma sobre o espaço das variáveis subexponenciais como o menor K_4 da propriedade (4).

Definição 3.2. *Seja X uma variável aleatória que satisfaz uma das propriedades equivalentes da Proposição 3.2. Então X é chamada de subexponencial. A norma subexponencial, denotada por $||X||_{\psi_1}$ é definida como*

$$||X||_{\psi_1} = \inf\{t > 0 : \mathbb{E}\left[\exp(|X|/t)\right] \leq 2\}.$$

Usando essa definição, obtemos um resultado de ligação entre variáveis subexponenciais e variáveis subgaussianas, enunciado a seguir.

Lema 3.1. *Uma variável aleatória X é subgaussiana se e somente se X^2 é subexponencial. Além disso,*

$$||X^2||_{\psi_1} = ||X||^2_{\psi_2}.$$

Demonstração. Da definição $||X||_{\psi_2}$ é o ínfimo dos números $t > 0$ tais que $\mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2$. Também temos que $||X^2||_{\psi_1}$ é o ínfimo dos números $k > 0$ tais que $\mathbb{E}\left[\exp(X^2/k)\right] \leq 2$. Portanto temos $t^2 = k$. □

Temos também a seguinte generalização do lema anterior.

Teorema 3.4. *Sejam X e Y variáveis aleatórias subgaussianas. Então XY é subexponencial. Além disso,*

$$||XY||_{\psi_1} \leq ||X||_{\psi_2}||Y||_{\psi_2}.$$

Demonstração. Assumimos inicialmente que $||X||_{\psi_2} = ||Y||_{\psi_2} = 1$. Pela definição de $||\cdot||_{\psi_2}$, isso implica que $\mathbb{E}\left[\exp(X^2)\right] \leq 2$ e $\mathbb{E}\left[\exp(Y^2)\right] \leq 2$. Queremos demonstrar que $\mathbb{E}\left[\exp(|XY|)\right] \leq 2$. Para isso, note que, para quaisquer $a, b \in \mathbb{R}^n$,, temos $a^2/2+b^2/2 \geq ab$. Então

$$\begin{aligned}
\mathbb{E}\left[\exp(|XY|)\right] &\leq \mathbb{E}\left[\exp\left(\frac{X^2}{2}+\frac{Y^2}{2}\right)\right] \\
&= \mathbb{E}\left[e^{X^2/2}e^{Y^2/2}\right] \\
&\leq \frac{1}{2}\mathbb{E}\left[(\exp(X^2/2))^2+(\exp(Y^2/2))^2\right] \\
&= \frac{1}{2}\left(\mathbb{E}\left[e^{X^2}\right]+\mathbb{E}\left[e^{Y^2}\right]\right) \\
&\leq \frac{1}{2}(2+2) = 2.
\end{aligned}$$

Para o caso em que as normas das subgaussianas não são iguais a 1, basta dividir as variáveis pela norma e aplicar o resultado já demonstrado, pois

$$\left|\left|\frac{XY}{||X||_{\psi_2}||Y||_{\psi_2}}\right|\right|_{\psi_1} \leq \left|\left|\frac{X}{||X||_{\psi_2}}\right|\right|_{\psi_2}\left|\left|\frac{Y}{||Y||_{\psi_2}}\right|\right|_{\psi_2},$$

isto conclui a demonstração. □

3.3 Desigualdade de Bernstein

Nesta seção iremos demonstrar uma desigualdade de concentração para soma de variáveis subexponenciais. Como já foi dito, a desigualdade de Bernstein será útil para estudar a concentração da norma do vetor com coordenadas subgaussinas, pois mostra que a soma variaveis subexponenciais (subgaussianas ao quadrado) também satisfazem uma desigualdade de concentração.

Teorema 3.5 (Desigualdade de Bernstein)**.** *Sejam $X_1, \ldots, X_n$ variáveis aleatórias subexponenciais independentes com média* 0*. Então, para alguma constante $c > 0$ e todo $t \geq 0$,*

temos

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^{n}||X_i||_{\psi_1}}, \frac{t}{\max_i ||X_i||_{\psi_1}}\right)\right).$$

Demonstração. Seja λ tal que $|\lambda| \leq c/\max_i ||X_i||_{\psi_1}$, então, pela afirmação (5) da Proposição 3.2, temos que

$$\mathbb{E}\left[e^{\lambda X_i}\right] \leq \exp(C\lambda^2 ||X_i||_{\psi_1}).$$

Portanto, pela desigualdade de Markov temos

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp(-\lambda t + C\lambda^2\sigma^2),$$

em que $\sigma^2 = \sum_{i=1}^{n} ||X_i||_{\psi_1}$. Para minimizar em λ, sob a restrição inicial basta escolher

$$\lambda_t = \min\left(\frac{t}{2C\sigma^2}, \frac{t}{\max_i ||X_i||_{\psi_1}}\right),$$

isto gera o limitante

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(-min\left(\frac{t^2}{4C\sigma^2}, \frac{ct}{2\max_i ||X_i||_{\psi_1}}\right)\right).$$

Para provar a desigualdade com o módulo, basta aplicar o mesmo procedimento com $-X_i$ em vez de X_i. □

3.4 Concentração da norma

Assuma que um vetor aleatório $X = (X_1, \ldots, X_n)$ possui coordenadas com média 0 e variância 1. Qual é a norma euclidiana média de X? Note que

$$\mathbb{E}\left[||X||_2^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] = n,$$

com isso, concluímos que podemos esperar que $||X||_2 \approx \sqrt{n}$. No teorema seguinte iremos mostrar que, com alta probabilidade $||X||_2$ está muito próximo de $\sqrt{n}$.

Teorema 3.6. *Seja $X = (X_1, \ldots, X_n)$ um vetor aleatório com coordenadas independentes, média 0 e subgaussianas satisfazendo $\mathbb{E}[X_i^2] = 1$. Então, para todo $t > 0$, temos*

que

$$\mathbb{P}\left(\left|||X||_2 - \sqrt{n}\right| \geq t\right) \leq 2\exp(-ct^2/K^4).$$

Demonstração. Note que

$$||X||_2^2 - n = \sum_{i=1}^{n}(X_i^2 - 1).$$

Uma vez que X_i é subgaussiana, temos que $X_i^2 - 1$ é subexponencial. Portanto, aplicando a desigualdade de Bernstein 3.5, temos , para todo $u \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}||X_i||_2^2 - 1\right| \geq u\right) \leq 2\exp\left(-c_1 \min\left(\frac{u^2 n}{K^4}, \frac{un}{K^2}\right)\right) \leq 2\exp\left(\frac{-c_1}{K^4}\min(u, u^2)\right).$$

Para o caso de $||X||_2$, usamos o seguinte fato:

$$|z - 1| \geq \delta \text{ implica que } |z^2 - 1| \geq \max(\delta, \delta^2),$$

ou seja, para todo $\delta \geq 0$, temos que

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}||X||_2 - 1\right| \geq \delta\right) &\leq \mathbb{P}\left(\left|\frac{1}{n}||X||_2^2 - 1\right| \geq \max(\delta^2, \delta)\right) \\ &\leq 2\exp\left(\frac{-cn}{K^4}\delta^2\right).\end{aligned}$$

Fazendo a mudança de variável $t = \delta\sqrt{n}$ obtemos o resultado desejado. □

Do Teorema 3.6, concluímos que um vetor aleatório cujas coordenadas são subgaussianas tende a se concentrar em uma fina casca ao redor da bola de raio $\sqrt{n}$.

Para visualizar esse comportamento, realizamos uma simulação em que foram gerados 10000 vetores aleatórios em dimensão $n = 1000$. Cada vetor foi amostrado com coordenadas independentes e distribuição $\mathcal{N}(0, 1)$. Em seguida, calculamos as normas desses vetores dividas por $\sqrt{n}$ e plotamos as frequências com que cada valor ocorreu. Os resultados podem ser observados na Figura 3.1.

Observe que a maior parte das amostras apresenta norma muito próxima de 1, evidenciando que, em alta dimensão, a distribuição da norma euclidiana dos vetores se concentra em $\sqrt{n}$.

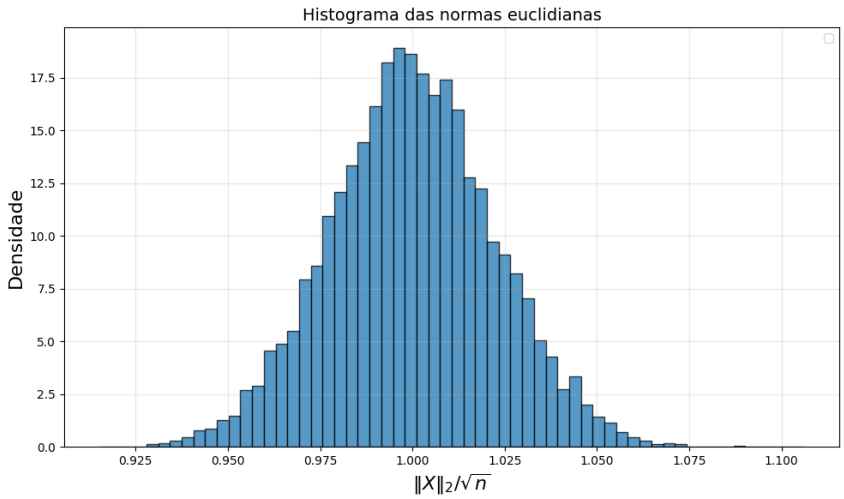

Figura 3.1: Histograma das normas para dimensão $n = 1000$.

3.5 Vetores isotrópicos e produtos ortogonais

Em um curso mais introdutório de probabilidade, um dos conceitos mais importantes é o de variância de uma distribuição, assim como a covariância, uma medida de variação entre duas variáveis diferentes.

Para o caso de um vetor de variáveis aleatórias, a covariância corresponde a uma matriz cujas entradas correspondem à covariância entre os pares de coordenadas do vetor. Mais precisamente, para um vetor aleatório X com média μ, a covariância de X é dada por

$$\Sigma_X = \mathbb{E}\left[XX^T\right] - \mu\mu^T.$$

A matriz Σ_X é uma matriz simétrica e semidefinida positiva. Portanto, de acordo com o teorema espectral (Axler, 2015), pode ser expressa por

$$\Sigma_X = \sum_i^n \lambda_i v_i v_i^T,$$

em que λ_i são os autovalores de Σ_X e v_i são os autovetores.

Ao lidar com vetores aleatórios, também é importante o conceito de isotropia, que generaliza a noção de variância unitária para variáveis aleatórias.

Definição 3.3 (Vetores aleatórios isotrópicos)**.** *Um vetor aleatório $X \in \mathbb{R}^n$ é chamado de isotrópico se*

$$\mathbb{E}\left[XX^T\right] = I_n,$$

em que I_n é a matriz identidade de dimensão $n \in \mathbb{N}$.

A partir dessa definição, é possível deduzir uma outra maneira de caracterizar vetores isotrópicos e o **produto interno** entre eles.

Definição 3.4 (Produto interno)**.** *Para todo $n \in \mathbb{N}$ e vetores $x, y \in \mathbb{R}^n$ de coordenadas $x = (x_1, \ldots, x_n)^T$ e $y = (y_1, \ldots, y_n)^T$, o* ***produto interno*** *entre x e y é definido como*

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i.$$

Lema 3.2 (Caracterização de isotropia)**.** *Um vetor aleatório $X \in \mathbb{R}^n$ é isotrópico se e somente se*

$$\mathbb{E}\left[\langle X, a\rangle^2\right] = ||a||_2^2, \text{ para todo } a \in \mathbb{R}^n.$$

Demonstração. Duas matrizes simétricas A e B são iguais se e somente se $a^T A a = a^T B a$ para todo $a \in \mathbb{R}^n$. Para provar isso, note que

$$a^T A a = a^T B a \iff a^T (A - B) a = 0.$$

Uma vez que a matriz $A - B$ é simétrica, podemos fazer sua decomposição espectral por $A - B = U^T \Lambda U$ em que U é uma matriz ortogonal e Λ é uma matriz diagonal. Assim,

$$a^T U^T \Lambda U a = 0.$$

Uma vez que U é uma matriz ortogonal, temos que $z = Ua$ será meramente uma rotação do vetor a, assim, podemos dizer que, para todo $z \in \mathbb{R}^n$,

$$z^T \Lambda z = 0,$$

ou seja, matriz dos autovalores Λ será a matriz nula, o que só ocorre se $A - B = 0$. Dessa forma, suponha então que X é um vetor isotrópico, então, para todo $a \in \mathbb{R}^n$, temos que

$$\begin{aligned}
\mathbb{E}\left[\langle X, a\rangle^2\right] &= \mathbb{E}\left[a^T X X^T a\right] \\
&= a^T \mathbb{E}\left[X X^T\right] a \\
&= a^T I_n a \qquad \text{(isotropia)} \\
&= ||a||_2^2.
\end{aligned}$$

Reciprocamente suponha que X é um vetor aleatório tal que $\mathbb{E}\left[\langle X, a\rangle^2\right] = ||a||_2^2$ para todo

$a \in \mathbb{R}^n$, então, do desenvolvimento acima, temos que

$$a^T \mathbb{E}\left[XX^T\right] a = a^T I_n a,$$

o que implica que $\mathbb{E}\left[XX^T\right] = I_n$, ou seja, X será isotrópico. □

Com a caracterização de isotropia dada acima, chegamos no próximo lema, que nos dá uma maneira de caracterizar o produto interno de dois vetores isotrópicos independentes como $\mathbb{E}\left[\langle X, Y\rangle^2\right] = n$.

Lema 3.3. *Sejam X e Y vetores isotrópicos em $\mathbb{R}^n$. Então*

$$\mathbb{E}\left[||X||_2^2\right] = n \text{ e } \mathbb{E}\left[\langle X, Y\rangle^2\right] = n.$$

Demonstração. Note que $X^T X$ é um escalar, portanto, da definição e propriedades do traço: $X^T X = tr(X^T X) = tr(XX^T)$. Assim

$$\mathbb{E}\left[||X||_2^2\right] = \mathbb{E}\left[X^T X\right] = \mathbb{E}\left[tr(XX^T)\right] = n,$$

em que a última igualdade decorre da isotropia de X. Para provar a segunda afirmação, aplicamos o Lema 3.2 condicionando a esperança em $x = Y$. Ou seja,

$$\begin{aligned}
\mathbb{E}\left[\langle X, Y\rangle^2\right] &= \mathbb{E}_Y \mathbb{E}_X\left[\langle X, Y\rangle^2 | Y\right] \\
&= \mathbb{E}_Y\left[||Y||_2^2\right] && \text{(Lema 3.3)} \\
&= n && \text{(da primeira parte)}
\end{aligned}$$

□

Uma aplicação interessante deste resultado é notar que, sendo X e Y dois vetores isotrópicos independentes, podemos padronizá-los fazendo

$$\bar{X} = \frac{X}{||X||_2} \text{ e } \bar{Y} = \frac{Y}{||Y||_2}.$$

A partir do Lema 3.3 e do Teorema 3.6, temos que $||X||_2 = \mathcal{O}(n)$ e $||Y||_2 = \mathcal{O}(n)$ e $\langle X, Y\rangle = \mathcal{O}(\sqrt{n})$, o que implica que

$$|\langle \bar{X}, \bar{Y}\rangle| \leq \frac{1}{\sqrt{n}},$$

ou seja, os vetores se tornam aproximadamente ortogonais à medida que n cresce. Como explicamos no começo do capítulo, uma razão intuitiva para isso é que em altas dimensões temos muito mais espaço, fazendo com que direções aleatórias se tornem cada vez mais distantes uma da outra, ou seja, quase ortogonais.

Como ilustração, consideremos, na Figura 3.2, uma simulação em que geramos 1000 pares de vetores aleatórios $\mathcal{N}(0,1)$ para diferentes dimensões e calculamos a distribuição do ângulo entre eles, o qual está diretamente relacionado ao produto interno. Podemos observar que, conforme n cresce, os ângulos gerados se concentram em torno de $\pi/2 \approx 1{,}57$ (ou 90°), em concordância com a discussão anterior.

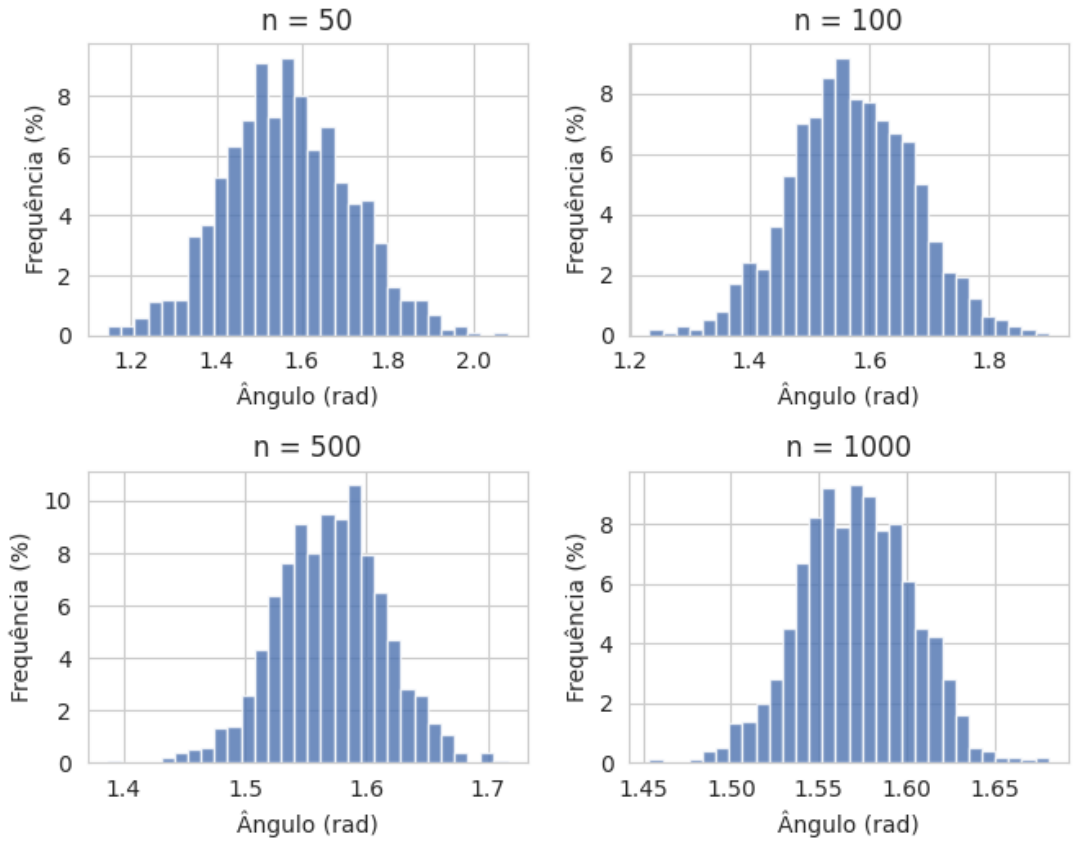

Figura 3.2: Histograma de ângulos (em radianos) entre dois vetores $\mathcal{N}(0,1)$ em $\mathbb{R}^n$ para $n = 50$, $n = 100$, $n = 500$ e $n = 1000$.

Uma consequência interessante deste fato é que que se tomarmos a distância entre dois vetores normalizados X e Y, temos que

$$||X - Y||_2^2 = ||X||_2^2 + ||Y||_2^2 - 2\langle X, Y\rangle \approx ||X||_2^2 + ||Y||_2^2 \approx 2,$$

isto é, os vetores estarão aproximadamente em posições diametralmente opostas na es-

fera. Assim, métodos de *machine learning* que dependem do cálculo de distâncias são negativamente afetados por este efeito (Bach, 2024).

3.6 Aplicação: Lema de Johnson-Lindestrauss

Como vimos na seção anterior, métodos de *machine learning* que usam como referência a distância entre os vetores não funcionam bem em altas dimensões. Uma maneira de contornar este problema é através de um procedimento de redução de dimensionalidade, que de alguma forma consiga preservar as distâncias entre os vetores.

Mais precisamente, suponha que tenhamos um conjunto de dados $X_1, \ldots, X_m \in \mathbb{R}^n$ em que $n \in \mathbb{N}$ é grande. Nosso objetivo é encontrar uma função $f : \mathbb{R}^n \to \mathbb{R}^d$ que satisfaça duas características: a primeira é que d seja consideravelmente menor que n. A segunda é que, para todo par X_i e X_j com $1 \leq i, j \leq m$ distintos tenhamos

$$||X_i - X_j||_2 \approx ||f(X_i) - f(X_j)||_2.$$

O Lema de Johnson-Lindestrauss, principal resultado desta seção, afirma que, não só tal função existe, como podemos usar uma construção aleatória para produzi-la com alta probabilidade.

A construção em questão faz uso de uma matriz aleatória. As matrizes aleatórias serão o tema de estudo central do próximo capítulo, em que explicaremos suas propriedades com mais detalhes. Por enquanto, basta saber que matrizes aleatórias são matrizes cujas entradas são variáveis aleatórias.

Lema 3.4 (Lema de Johnson-Lindestrauss)**.** *Sejam m, n e d inteiros positivos e $X_1, \ldots, X_m \in \mathbb{R}^n$ uma sequência de vetores aleatórios independentes. Considere $Z \in \mathbb{R}^{d \times n}$ uma matriz cujas entradas Z_{ij} com $1 \leq i \leq d$ e $1 \leq j \leq n$ são variáveis aleatórias $\mathcal{N}(0,1)$ independentes. Defina a função $f : \mathbb{R}^n \to \mathbb{R}^d$ por*

$$f(X) = \frac{ZX}{\sqrt{d}}.$$

Então, para todo $\varepsilon \in (0,1)$, existe um $\delta \in (0,1)$ tal que, para todo par X_i e X_j distintos, existe uma constante $c > 0$ tal que, com probabilidade pelo menos $1 - \delta$,

$$(1-\varepsilon)||X_i - X_j||_2 \leq ||f(X_i) - f(X_j)||_2 \leq (1+\varepsilon)||X_i - X_j||_2,$$

desde que

$$d \geq \frac{c \log(n/\delta)}{\varepsilon^2}.$$

Demonstração. As variáveis Z_{ij} são normais com média 0 e variância 1. Do Teorema 3.6, isto implica que cada linha da matriz, denotada por Z_i, satisfaz, para alguma constante $c > 0$ e todo $t > 0$,

$$\mathbb{P}\left(\left|\frac{1}{d}\sum_{j=1}^{d} Z_{ij} - 1\right| \geq t\right) \leq 2\exp(-cdt^2). \tag{3.1}$$

Fixe agora um par (X_j, X_k) e considere,

$$\begin{aligned}\frac{||f(X_j) - f(X_K)||_2^2}{||X_j - X_k||_2^2} &= \left|\left|\frac{Z(X_j - X_k)}{\sqrt{d}||X_j - X_k||_2}\right|\right|_2^2 \\ &= \frac{1}{d}\sum_{i=1}^{d}\left\langle Z_i, \frac{X_j - X_k}{||X_j - X_k||_2}\right\rangle^2.\end{aligned}$$

Cada termo da última soma acima é uma variável subgaussiana, portanto, podemos aplicar o limitante em 3.1, obtendo

$$\mathbb{P}\left(\left|\frac{||f(X_j) - f(X_k)||_2^2}{||X_j - X_k||_2^2} - 1\right| \geq \varepsilon\right) \leq 2\exp(-cd\varepsilon^2).$$

O evento da probabilidade acima representa a falha da afirmação do lema apenas para o par (X_j, X_k). Para estimar a probabilidade de que algum dos pares possíveis da nossa sequência contrarie essa afirmação, podemos usar o limitante da união dos eventos para cada par:

$$\mathbb{P}\,(\text{falha}) \leq \binom{m}{2} 2\exp(-cd\varepsilon^2).$$

Pode-se verificar que, se

$$d \geq \frac{c \log(m/\delta)}{\varepsilon^2},$$

então essa probabilidade é no máximo δ, como queríamos demonstrar. □

Para ilustrar empiricamente o Lema de Johnson–Lindenstrauss, realizamos uma simulação numérica baseada em uma única projeção aleatória. Foram gerados $m = 10{,}000$ vetores $X_1, \ldots, X_m \in \mathbb{R}^n$ com $n = 1000$, cujas coordenadas seguem uma distribuição normal padrão.

Em seguida, todos os vetores foram projetados em um subespaço de dimensão $d = 50$ utilizando uma única matriz aleatória $Z \in \mathbb{R}^{d \times n}$ com entradas independentes $\mathcal{N}(0, 1)$, de

acordo com

$$f(X) = \frac{ZX}{\sqrt{d}}.$$

Para analisar a preservação das distâncias, calculamos a razão

$$R_{ij} = \frac{\|f(X_i) - f(X_j)\|_2}{\|X_i - X_j\|_2}$$

para uma amostra de pares de vetores (X_i, X_j) selecionados aleatoriamente entre os 10.000 vetores gerados. O histograma apresentado na Figura mostra a distribuição empírica dessas razões em porcentagem.

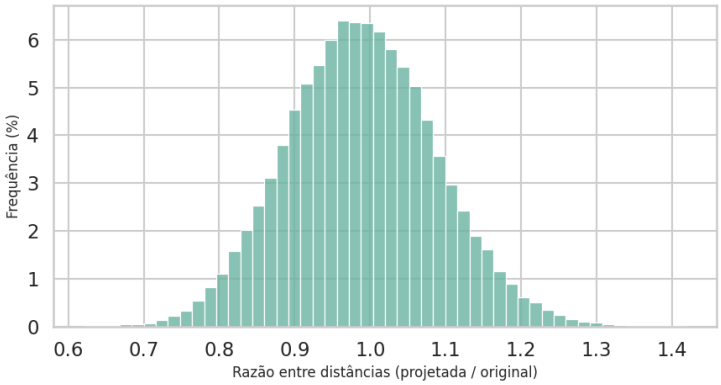

Figura 3.3: Histograma das razões R_{ij}

Observa-se que os valores de R_{ij} se concentram fortemente em torno de 1, indicando que a distância entre os vetores é aproximadamente preservada após a projeção. Dessa forma, a simulação evidencia de maneira prática que, mesmo em conjuntos de dados muito grandes e de alta dimensão, uma única matriz aleatória consegue preservar a geometria do conjunto de pontos.

Capítulo 4

Matrizes aleatórias e detecção de comunidades em redes

Neste capítulo estudaremos o conceito de matrizes aleatórias e desigualdades de concentração para seus valores singulares. O objetivo será a demonstração do funcionamento de um algoritmo para o problema de detecção de comunidades em redes.

Na Seção 4.1, traremos alguns conceitos importantes de álgebra linear e mostramos como o maior valor singular pode ser entendido como uma norma no espaço das matrizes e como elas se relacionam com a separação do espectro das matrizes, através da desigualdade de Davis-Kahan. A Seção 4.2 desenvolve o conceito de espaços métricos, que será usado para estudar a concentração de valores singulares na Seção 4.3.

Por fim, na Seção 4.4, mostraremos como os valores singulares de matrizes aleatórias estudadas nas seções anteriores podem ser aplicadas no estudo das redes aleatórias, mais especificamente, no problema de detecção de comunidades.

4.1 Revisão sobre matrizes e desigualdade de Davis-Kahan

Nesta seção apresentamos uma breve recapitulação de alguns tópicos fundamentais de álgebra linear, com ênfase na decomposição em valores singulares. Como referência básica para esta revisão, indicamos os textos clássicos de Axler (2015) e Horn e Johnson (2012).

No capítulo anterior, fizemos uso da decomposição espectral, que é válida exclusivamente para matrizes reais simétricas. A decomposição em valores singulares é uma

generalização deste conceito para todas as matrizes, isto é, seja A uma matriz real com m linhas e n colunas, então existem $s_1, \ldots, s_r \geq 0$ com $r = min(m, n)$ tais que

$$A = \sum_{i=1}^{r} s_i u_i v_i^T,$$

em que $(u_i)_{i=1}^r \in \mathbb{R}^m$ e $(v_i)_{i=1}^r \in \mathbb{R}^n$ são sequências de vetores ortogonais. Os valores s_i são chamados de valores singulares de A, os vetores u_i de autovetores à esquerda e os vetores v_i de autovetores à direita. No caso de matrizes simétricas, os valores singulares correspondem aos autovalores e os autovetores à direita e à esquerda são iguais.

Um conceito importante que abordaremos nas seções seguintes será o de norma do operador, definido a seguir.

Definição 4.1 (Norma do operador)**.** *A norma do operador de uma matriz A de dimensão $m \times n$ é definida como o menor número K tal que*

$$||Ax||_2 \leq K||x||_2, \text{ para todo } x \in \mathbb{R}^n.$$

A norma do operador equivale ao maior valor singular de A e, portanto, mede a maior dilatação que a norma de um vetor submetido à transformação A pode sofrer.

O principal teorema desta seção é a desigualdade de Davis-Kahan, que nos dá uma forma de aproximar o ângulo entre dois autovetores de matrizes simétricas usando a norma do operador. Começamos com uma versão mais forte da desigualdade, que se refere a projeções espectrais. Para isso, introduziremos o conceito de conjuntos δ-separados, do qual a desigualdade depende.

Definição 4.2. *Dois conjuntos I e J de $\mathbb{R}$ são denominados δ-separados quando, dado $\delta > 0$, para todo $x \in I$ e todo $y \in J$ temos que*

$$|x - y| \geq \delta.$$

Teorema 4.1 (Desigualdade de Davis-Kahan para projeções espectrais)**.** *Considere duas matrizes simétricas A e B de dimensões $n \times n$ e decomposições espectrais $A = \sum_{i=1}^n \lambda_i u_i u_i^T$ e $B = \sum_{j=1}^n \mu_j v_j v_j^T$. Sejam I e J dois conjuntos δ-separados subconjuntos de $\mathbb{R}$, em que*

I é um intervalo. Então as projeções espectrais

$$P = \sum_{\lambda_i \in I} u_i u_i^T, \quad Q = \sum_{\mu_j \in J} v_j v_j^T \text{ satisfazem } ||QP|| \leq \frac{||A - B||}{\delta}.$$

Demonstração. Assuma que o intervalo I é finito e fechado. Se somarmos o mesmo múltiplo de identidade a A e B, podemos centrar I como $[-r, r]$, retornado $|\lambda_i| \leq r$ para $\lambda_i \in I$ e $|\mu_j| \geq r + \delta$ para $\mu_j \in J$. Definindo a matriz $H = A - B$, temos

$$||H|| = ||QHP|| = ||QBP - QAP|| \geq ||QBP|| - ||QAP||. \tag{4.1}$$

A projeção espectral Q comuta com B, então

$$||QBP|| = ||BQP|| \leq (r + \delta)||QP||. \tag{4.2}$$

A última desigualdade decorre do fato de que a imagem de Q é uma combinação linear de autovetores v_j, quando aplicamos a matriz B nessa imagem enviamos cada autovetor v_j para $\mu_j v_j$, portanto reescalando-o por pelo menos $r + \delta$. Além disso, $AP = PAP$, então

$$||QAP|| = ||QPAP|| \leq ||QP|| \cdot ||AP|| \leq r||QP||, \tag{4.3}$$

porque $||AP|| = \max_{i \in I} \lambda_i \leq r$. Substituindo 4.2 e 4.3 em 4.1, obtemos que $||H|| \geq \delta||QP||$, como queríamos demonstrar. □

Na versão da desigualdade de Davis-Kahan que provaremos a seguir faremos também uso de um resultado preliminar chamado de desigualdade de Weyl (Horn e Johnson, 2012), cuja função é limitar a diferença entre valores singulares de matrizes usando a norma do operador. A relação entre os resultados vêm do fato de que, enquanto a desigualdade de Weyl mostra o que ocorre com a matriz a partir de perturbações em seus autovalores, a desigualdade de Davis-Kahan explica como os autovetores são afetados.

Lema 4.1 (Desigualdade de Weyl)**.** *Os k-ésimos maiores valores singulares das matrizes A e B (denotados por $s_k(A)$ e $s_k(B)$) de dimensões $m \times n$ satisfazem*

$$|s_k(A) - s_k(B)| \leq ||A - B||.$$

Teorema 4.2 (Desigualdade de Davis-Kahan)**.** *Considere duas matrizes simétricas A e*

B com decomposições espectrais $A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$ e $B = \sum_{j=1}^{n} \mu_j v_j v_j^T$ onde os autovalores estão ordenados de forma decrescente. Assuma que o k-ésimo auto-valor de A é δ-separado do resto:

$$\min_{i:i\neq k} |\lambda_i - \lambda_k| = \delta > 0,$$

então o seno do ângulo entre os autovetores u_k e v_k satisfaz

$$\sin \angle u_k, v_k \leq \frac{2||A-B||}{\delta}.$$

Demonstração. Pode-se assumir que $\varepsilon := ||A - B|| \leq \delta/2$, pois caso contrário a solução é trivial. Pela desigualdade de Weyl 4.1, $|\lambda_j - \mu_j| \leq \varepsilon$ para cada j, então

$$\min_{j:j\neq k} |\lambda_j - \mu_k| \geq \min_{j:j\neq k} |\lambda_k - \lambda_j| - \varepsilon = \delta - \varepsilon \geq \delta/2.$$

Aplicando o Teorema 4.1 para os conjuntos $I = \{\lambda_k\}$ e $J = \{\mu_j : j \neq k\}$ para obter $||QP|| \leq 2\varepsilon/\delta$. Por fim, note que P e $I_n - Q$ são projeções ortogonais nas direções de u_k e v_k, então

$$||QP|| = ||Qu_k||_2 = \sin \angle u_k, v_k.$$

□

4.2 Introdução aos espaços métricos

Na estatística existem muitos exemplos de processos estocásticos em que as variáveis aleatórias são indexados por um conjunto contínuo (Wainwright, 2019b). Para ajudar a entender fenômenos desse tipo começamos esta seção com a definição de um espaço métrico.

Definição 4.3 (Espaços métricos)**.** *Um espaço métrico é um conjunto não-vazio T em que podemos definir uma função $\rho : T \times T \to \mathbb{R}$ que satisfaz as seguintes propriedades, para quaisquer $\theta_1, \theta_2, \theta_3 \in T$:*

- *Não negatividade: $\rho(\theta_1, \theta_2) \geq 0$ com igualdade somente quando $\theta_1 = \theta_2$.*
- *Simetria: $\rho(\theta_1, \theta_2) = \rho(\theta_2, \theta_1)$.*
- *Desigualdade triangular: $\rho(\theta_1, \theta_3) \leq \rho(\theta_1, \theta_2) + \rho(\theta_2, \theta_3)$.*

Exemplos de espaços métricos são o conjunto $\mathbb{R}^d$ com a distância euclidiana e o conjunto $\{0, 1\}$ com a métrica reescalada de Hamming:

$$\rho_H(\theta, \theta') = \frac{1}{d} \sum_{i=1}^{d} \mathbb{I}(\theta_j \neq \theta'_j). \tag{4.4}$$

Uma maneira natural de medir o tamanho de um espaço métrico é através do número de bolas de raio δ necessário para cobri-lo. Para isso, definimos formalmente a seguir o conceito de cobertura-δ e de número de cobertura.

Definição 4.4 (Número de cobertura)**.** *Seja T um espaço métrico com métrica ρ. Uma cobertura-δ de um conjunto $K \subset T$ é um conjunto $\{\theta_1, \ldots, \theta_N\}$ de elementos de K tais que, para todo $\theta \in K$, $\rho(\theta, \theta_i) \leq \delta$ para algum $i \in 1, \ldots, N$. O número de cobertura $N(K, \rho, \delta)$ se refere ao número de elementos da menor cobertura-δ.*

O número de cobertura pode ser visto como uma medida de tamanho de um conjunto em um espaço métrico. O valor $N(\delta)$ é uma função não-crescente de δ e diverge quando $\delta \to 0^+$.

Ainda relacionado com a noção de cobertura, temos a ideia de empacotamento.

Definição 4.5 (Conjuntos δ - separados)**.** *Um subconjunto $\mathcal{N}$ de um espaço métrico T com métrica ρ é dito ser δ-separado se, para quaisquer x e y distintos em $\mathcal{N}$, temos que $\rho(x, y) > \delta$. O número de empacotamento de um dado conjunto $K \subset T$ é definido como o número de elementos do maior subconjunto δ-separado de K, escrito como $\mathcal{P}(K, \rho, \delta)$.*

Uma intuição para a definição acima é de que, dado um conjunto K, e um subconjunto δ - separado $\mathcal{N} \subset K$, sempre é possível colocar, ou empacotar, $\mathcal{P}(K, \rho, \delta)$ bolas de raio $\delta/2$ disjuntas centradas nos elementos de $\mathcal{N}$ no conjunto K. Isto porque se algum par de bolas fosse disjunto, haveria pelo menos um elemento $z \in T$ e $x, y \in \mathcal{N}$ tais que $\rho(x, z) \leq \delta/2$ e $\rho(y, z) \leq \delta/2$, o que pela desigualdade triangular retornaria $\rho(x, y) \leq \delta$, contradizendo a hipótese de que $\mathcal{N}$ é um conjunto δ-separado.

Podemos ligar os conceitos de cobertura e empacotamento usando o seguinte lema.

Lema 4.2 (Equivalência dos números de empacotamento e cobertura)**.** *Para qualquer conjunto $K \subset T$ e qualquer $\delta > 0$, temos que*

$$\mathcal{P}(K, \rho, 2\delta) \leq N(K, \rho, \delta) \leq \mathcal{P}(K, \rho, \delta).$$

Demonstração. No limitante superior, demonstraremos que, se $\mathcal{N}$ é um conjunto δ-separado maximal de K, então $\mathcal{N}$ é uma δ-cobertura de K. É necessário então que para todo elemento $x \in K$, exista algum $y \in \mathcal{N}$ tal que $\rho(x, y) \leq \delta$. Se $x \in \mathcal{N}$, então isto é obviamente verdade. Suponha agora que x não está em $\mathcal{N}$, se para todo $y \in \mathcal{N}$ tivéssemos $\rho(x, y) > \delta$, a maximalidade do conjunto $\mathcal{N}$ seria violada, ou seja, deve haver algum $y \in \mathcal{N}$ tal que $\rho(x, y) \leq \delta$.

No limitante inferior tome $\mathcal{P}$ um conjunto δ-separado e $\mathcal{N}$ uma cobertura-δ de K. Então qualquer ponto $x \in \mathcal{P}$ está em alguma bola fechada de raio δ centrada algum ponto $y \in \mathcal{N}$. Uma vez que uma bola fechada de raio δ não pode conter dois elementos 2δ-separados, cada bola deve conter no máximo um elemento de $\mathcal{P}$. Isto implica que $|\mathcal{P}| \leq |\mathcal{N}|$ e, como os conjuntos são arbitrários, o limitante inferior segue. □

4.2.1 Número de cobertura e volume em $\mathbb{R}^n$

De conhecimento dos resultados da seção anterior, seguimos agora para o exemplo mais importante de espaço métrico, $\mathbb{R}^n$, munido da distância euclidiana como métrica. Para facilitar a notação, consideraremos $N(K, \delta)$ o número de cobertura δ do conjunto $K \subset \mathbb{R}^n$ sob a distância euclidiana.

Gostaríamos de relacionar o número de cobertura com a medida mais usual de tamanho de um conjunto para este caso, o volume. Não existe uma equivalência total entre as duas pois podem haver conjuntos que possuem volume zero e número de cobertura positiva, por exemplo, um quadrado no plano $\mathbb{R}^2$. No entanto, podemos ter uma equivalência parcial, ainda útil. Essa equivalência depende de um conceito chamado de soma de Minkowski, definido abaixo.

Definição 4.6 (Soma de Minkowski)**.** *Sejam A e B dois subconjuntos de $\mathbb{R}^n$, a soma de Minkowski $A + B$ é definida como*

$$A + B = \{a + b : a \in A, b \in B\}.$$

Proposição 4.1. *Seja K um subconjunto de $\mathbb{R}^n$ e $\delta > 0$. Então*

$$\frac{Vol(K)}{Vol(\delta B_2^n)} \leq N(K, \delta) \leq \mathcal{P}(K, \delta) \leq \frac{Vol(K + (\delta/2)B_2^n)}{Vol(\delta/2)B_2^n},$$

em que B_2^n é a bola euclidiana de raio 1, ou seja, δB_2^n é a bola euclidiana de raio δ.

Demonstração. A segunda desigualdade já foi demonstrada no Lema 4.2. Para a primeira desigualdade, note que K pode ser coberto por $N(K,\delta)$ bolas euclidianas de raio δ, ou seja,

$$N(K,\delta)Vol(\delta B_2^n) \geq Vol(K).$$

Para a terceira desigualdade, note que podemos encontrar $\mathcal{P}(K,\delta)$ bolas de raio $\delta/2$ e centros x_i em K disjuntas. Mesmo que essas bolas não estejam inteiramente contidas em K, elas estarão contidas no conjunto $K + (\delta/2)B_2^n$, isto é,

$$\mathcal{P}(K,\delta)Vol((\delta/2)B_2^n) \leq Vol(K + (\delta/2)B_2^n).$$

□

Como consequência dos limitantes acima, é possível concluir que os números de cobertura, em geral, crescem de maneira exponencial com a dimensão n, como veremos no seguinte resultado.

Proposição 4.2. *Os número de cobertura da bola euclidiana B_2^n satisfazem, para todo $\delta > 0$:*

$$\left(\frac{1}{\delta}\right)^n \leq N(B_2^n,\delta) \leq \left(\frac{2}{\delta}+1\right)^n.$$

Demonstração. O volume da bola δB_2^n em $\mathbb{R}^n$ escala como $Vol(\delta B_2^n) = \delta^n Vol(B_2^n)$, o que gera imediatamente o limitante inferior. Para o limitante superior temos, usando 4.1

$$N(K,\delta) \leq \frac{(1+\delta/2)^n Vol(B_2^n)}{(\delta/2)^n Vol(B_2^n)} = \left(\frac{2}{\delta}+1\right)^n.$$

□

4.3 Valores singulares de matrizes subgaussianas

Nesta seção iremos introduzir nossos primeiros resultados sobre matrizes aleatórias, que dizem respeito a uma maneira de limitar, com alta probabilidade, a norma do operador de uma matriz A com dimensão $m \times n$. No processo de demonstração veremos a utilidade dos conceitos de cobertura apresentados anteriormente.

A norma do operador de uma matriz A, já definida anteriormente, pode ser escrita

como

$$||A|| = \max_{x \neq 0} \frac{||Ax||_2}{||x||_2} = \max_{x \in S^{n-1}} ||Ax||_2,$$

em que S^{n-1} é a esfera unitária em $\mathbb{R}^n$, isto é, o conjunto de todos os vetores $x \in \mathbb{R}^n$ tais que $||x||_2 = 1$ para todo $n \in \mathbb{N}$. A norma do operador é um valor que representa a dilatação máxima que um vetor sofre quando submetido à transformação linear A. O lema a seguir mostra que o máximo acima não precisa ser tomado em todos os vetores da esfera unitária, mas apenas a uma cobertura-δ desta.

Lema 4.3. *Seja A uma matriz $m \times n$ e $\delta \in [0, 1)$. Então, para qualquer cobertura-δ $\mathcal{N}$ da esfera S^{n-1}, temos que*

$$\sup_{x \in \mathcal{N}} ||Ax||_2 \leq ||A|| \leq \frac{1}{1-\delta} \sup_{x \in \mathcal{N}} ||Ax||_2.$$

Demonstração. O limitante inferior é claro já que $\mathcal{N} \subset S^{n-1}$. Para o limitante superior, seja x o vetor tal que $||Ax||_2 = ||A||$ e tome x_0 tal que $||x - x_0||_2 \leq \delta$. Usando a definição da norma do operador, temos que

$$\begin{aligned} ||Ax - Ax_0||_2 &= ||A(x - x_0)||_2 \\ &= \frac{||A(x - x_0)||_2}{||x - x_0||_2} ||x - x_0||_2 \\ &\leq ||A|| \cdot ||x - x_0||_2 \\ &\leq \delta ||A||. \end{aligned}$$

Usando a desigualdade triangular, obtemos

$$\begin{aligned} ||Ax_0||_2 &= ||Ax - (Ax - Ax_0)||_2 \\ &\geq ||Ax||_2 - ||Ax - Ax_0||_2 \\ &\geq ||A|| - \delta ||A|| = (1 - \delta)||A||, \end{aligned}$$

dividindo por $1 - \delta$ dos dois lados chegamos no limitante desejado. □

A partir do Lema 4.3, podemos deduzir que a norma do operador da matriz A pode ser encontrada maximizando uma forma quadrática:

$$||A|| = \sup_{x \in S^{n-1},\, y \in S^{m-1}} |\langle Ax, y \rangle|. \tag{4.5}$$

Com isso, podemos provar o primeiro resultado sobre matrizes aleatórias, que afirma que uma matriz com entradas subgaussianas de média zero possui norma do operador $||A|| \leq \sqrt{m} + \sqrt{n}$ com alta probabilidade.

Teorema 4.3. *Seja A uma matriz $m \times n$ aleatória com entradas subgaussianas, independentes com médias zero. Então, para todo $t > 0$, com probabilidade de pelo menos $1 - \exp(-t^2)$, temos que*

$$||A|| \leq CK(\sqrt{m} + \sqrt{n} + t),$$

onde $K = \max_{i,\, j} ||A||_{\psi_2}$ e $C > 0$ é uma constante.

Demonstração. Inicialmente, usaremos a equação 4.5 como nossa definição de norma do operador, em seguida, construiremos coberturas-δ para aproximar o espaço onde o máximo está sendo tomado. Assim, tome $\delta = 1/4$. Da Proposição 4.2, podemos encontrar redes-δ $\mathcal{N}$ e $\mathcal{M}$ para S^{n-1} e S^{m-1}, respectivamente, com cardinalidades $|\mathcal{N}| \leq 9^n$ e $\mathcal{M} \leq 9^m$.

Em seguida, fixe $x \in \mathcal{N}$ e $y \in \mathcal{M}$. A forma quadrática

$$\langle Ax, y \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} x_i y_j$$

é uma soma de variáveis subgaussianas. Do Teorema 3.1, sabemos que ela é subgaussiana e

$$\begin{aligned} ||\langle Ax, y \rangle||_{\psi_2}^2 &\leq C \sum_{i=1}^{n} \sum_{j=1}^{m} ||A_{ij} x_i y_j||_{\psi_2}^2 \\ &\leq CK^2 \sum_{i=1}^{n} \sum_{j=1}^{m} x_i^2 y_j^2 \\ &= CK^2 \left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{j=1}^{m} y_j^2 \right) = CK^2. \end{aligned}$$

Usando as propriedades de variáveis subgaussianas, podemos reescrever a desigualdade acima como $\mathbb{P}\left(|\langle Ax, y \rangle| \geq z\right) \leq 2\exp(-cz^2/K^2),\ z \geq 0$.

Por fim, podemos desfixar x e y e fazer um limitante da união de seguinte forma:

$$\mathbb{P}\left(\max_{x \in \mathcal{N},\, y \in \mathcal{M}} |\langle Ax, y \rangle| \geq z \right) \leq \sum_{x \in \mathcal{N},\, y \in \mathcal{M}} \mathbb{P}\left(|\langle Ax, y \rangle| \geq z\right) \leq 9^{n+m} 2\exp(-cz^2/K^2).$$

Tomando $z = CK(\sqrt{m} + \sqrt{n} + t)$, temos $z^2 \geq C^2K^2(n + m + t^2)$ e se a constante C for

grande o suficiente, podemos tomar o expoente $cz^2/K^2 \geq 3(n+m) + t^2$. Então

$$\mathbb{P}\left(||A|| \geq z\right) = \mathbb{P}\left(\max_{x \in \mathcal{N},\, ,y \in \mathcal{M}} |\langle Ax, y\rangle| \geq z\right) \leq 9^{n+m} 2\exp(-3(n+m) - t^2) \leq 2\exp(-t^2),$$

como queríamos demonstrar. □

Para ilustrar o Teorema 4.3, foi realizada uma simulação na qual geramos 100 matrizes com $m = 10$ linhas e variamos o número n de colunas. Cada uma das matrizes possui entradas normais $\mathcal{N}(0,1)$ e portanto possuem $K = \max_{ij} A_{ij} = \sqrt{8/3}$. Dessa forma, calcularemos os valores singulares de cada uma para estudar a distribuição, para diferentes valores de $t > 0$, de

$$R = \frac{||A||}{K(\sqrt{m} + \sqrt{n} + t)}.$$

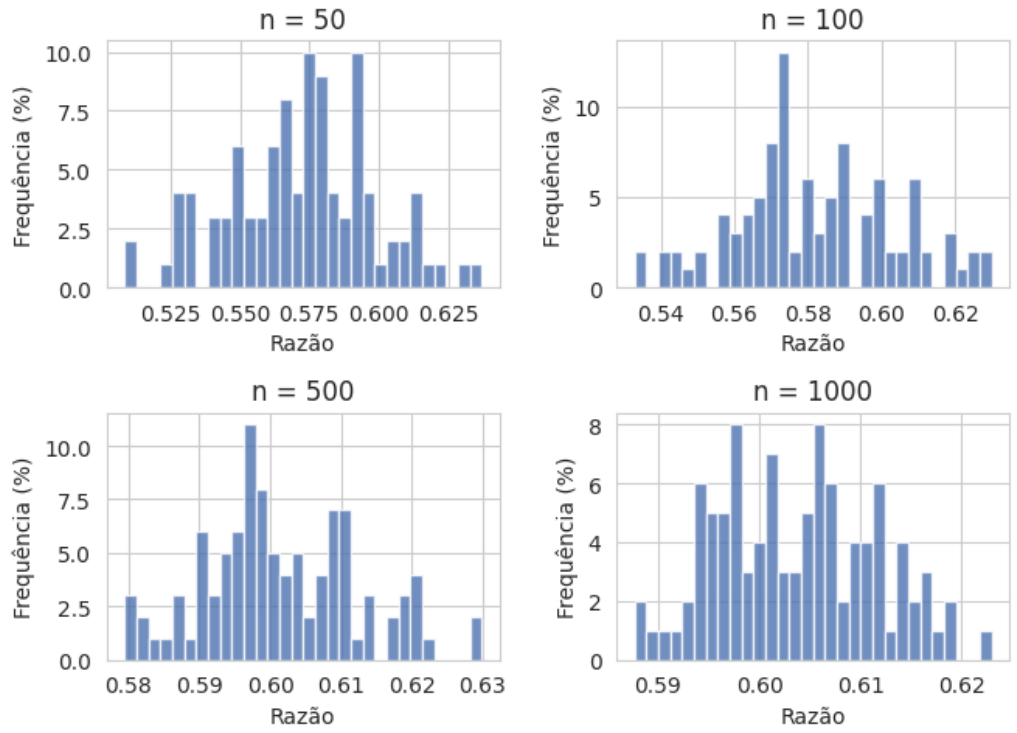

Figura 4.1: Histograma da razão R com $n = 50$, $n = 100$, $n = 500$ e $n = 1000$, mantendo $t = 0,01$.

Como podemos observar na Figura 4.1, a distribuição da razão R fica distribuída em um intervalo pequeno constante, nunca sendo da ordem das dimensões $\sqrt{m} + \sqrt{n}$.

Embora o Teorema 4.3 acima se aplique a um caso já bastante geral no espaço das matrizes, existe uma extensão útil dele para as matrizes simétricas, em que as entradas não são independentes.

Corolário 4.1. *Seja A uma matriz simétrica $n \times n$ com entradas subgaussianas e média zero, independentes na diagonal e acima dela. Então, para todo $t > 0$, temos que*

$$||A|| \leq CK(\sqrt{n} + t),$$

com probabilidade pelo menos $1 - 4\exp(-t^2)$ e $K = \max_{ij} ||A_{ij}||_{\psi_2}$

Demonstração. Podemos partir em A em uma soma de matrizes triangulares com A^+ sendo uma triangular contendo a diagonal de A e A^- uma triangular inferior com 0 na diagonal. Note que como as entradas serão subgaussianas independentes, podemos aplicar o Teorema 4.3 a cada uma das matrizes separadamente. Então, para todo $t > 0$, por um limitante de união, com probabilidade $1 - 4\exp(-t^2)$ temos que $||A^+|| \leq CK(\sqrt{n} + t)$ e $||A^-|| \leq CK(\sqrt{n}+t)$. Pelas propriedades da norma, temos que $||A|| \leq ||A^-|| + ||A^+||$. □

4.4 Aplicação: deteccção de comunidades em redes

O modelo de blocos estocásticos para um grafo aleatório é uma extensão do modelo clássico de Erdos-Renyi (Bollobás, 1998). No modelo de Erdos-Renyi clássico temos um grafo com n vértices, em que cada par de vértices pode ser ligado por uma aresta com probabilidade $p \in (0, 1)$ ou não ser ligado por uma aresta com probabilidade $1 - p$, de forma independente. A diferença é que aqui estamos dividindo o conjunto de n vértices em dois subconjuntos (comunidades) de $n/2$ vértices. Um par de vértices é conectado por uma aresta com probabilidade p se eles pertencem à mesma comunidade e, caso sejam de comunidades diferentes, serão ligado com probabilidade q. Denotaremos esse grafo aleatório por $G(n, p, q)$.

No caso especial em que p e q são iguais, temos o grafo aleatório de Erdos-Renyi, mas quando $p > q$, conexões acontecem com maior frequência dentro de uma comunidade que fora dela. Na prática, não conhecemos p e q e nem a estrutura real das comunidades. Portanto, a realização de um grafo $G(n, p, q)$ em que conhecemos seu vértices e as conexões entre eles, como determinamos quais são as duas comunidades do grafo?

Para estudar essa estrutura, podemos usar uma ferramenta chamada de matriz de adjacência do grafo, definida abaixo.

Definição 4.7 (Matriz de adjacência de um grafo)**.** *Seja G um grafo com n vértices, rotulados pelo conjunto $\{1, \ldots, n\}$. A matriz de adjacência de G é a matriz A de dimensão*

$n \times n$ cujas entradas A_{ij} são iguais a 1 se os vértices i e j distintos tem uma aresta entre si e 0 caso contrário. Se $i = j$, convenciona-se que $A_{ij} = 0$.

No caso do nosso modelo, por definição a matriz de adjacência será aleatória e suas entradas serão variáveis Bernoulli com parâmetro p ou q, a depender das entradas. Tratam-se, portanto, de variáveis subgaussianas e poderemos usar os resultados demonstrados anteriormente. Para isso, iremos escrever A como uma soma em uma parte determinística e uma aleatória da seguinte forma:

$$A = D + R,$$

em que $D = \mathbb{E}[A]$, que contém a parte informativa e R é uma matriz aleatória que pensaremos como se um ruído.

As entradas da matriz D são p ou q e podemos ordenar suas entradas de acordo com as comunidades dos vértices. Além disso, a matriz possui posto 2 e sua decomposição espectral retorna como autovalores $\lambda_1 = (p+q)n/2$ e $\lambda_2 = (p-q)n/2$, cujos respectivos autovetores são $u_1 = (1, 1, \ldots, 1)$ e $u_2 = (1, 1, \ldots, 1, -1, -1, \ldots, -1)$. O segundo autovetor u_2 contém toda a informação que precisamos para determinar as comunidades (vértices com coordenadas iguais pertencem à mesma comunidade). No entanto, não conhecemos a matriz D e não podemos acessar u_2 diretamente. Mas conhecemos a matriz de adjacência A real e podemo usá-la como uma estimativa de D. Note que o sinal de D é da ordem de n enquanto o sinal de R é da ordem de $\sqrt{n}$ com alta probabilidade, como demonstramos no Corolário 4.1.

Para demonstrar formalmente a afirmação de que, nesse contexto, a matriz A é de fato uma boa estimativa para D, iremos usar a desigualdade de Davis-Kahan. Começamos verificando o quão separado do resto do espectro de D está o seu segundo autovalor. Isto é:

$$\min(\lambda_2(D), \lambda_1(D) - \lambda_2(2)) = \min\left(\frac{p-q}{2}, q\right) n := \mu n.$$

A desigualdade de Davis-Kahan nos dá um limitante útil em termos de R e de δ para o seno do ângulo entre os autovetores unitários de A e D:

$$\sin \angle \bar{u_2}(D), \bar{u_2}(A) \leq \frac{2||R||}{\delta} \leq \frac{C}{\mu\sqrt{n}}.$$

Se o seno dos ângulos unitários é pequeno, então existe uma constante $\theta \in \{-1, 1\}$ tal

que

$$||\bar{u_2}(A) - \theta u_2(\bar{D})||_2 \leq C/\mu^2.$$

Uma vez que o vetor $u_2(D)$ possui norma $\sqrt{n}$, podemos multiplicar por $\sqrt{n}$ dos dois lados para obter

$$||u_2(A) - \theta u_2(D)||_2 \leq C/\mu^2,$$

isto é, a maior parte dos sinais dos dois autovetores coincidem.

Veremos agora o funcionamento deste algoritmo na prática, isto é, usando um banco de dados reais. A partir da matriz de adjacências e do segundo autovetor resultado de sua decomposição, iremos fazer a detecção de comunidades.

O banco de dados em questão é conhecido como *Zachary's Karate Club*, que foi obtido em um estudo realizado pelo antropólogo Wayne Zachary (Zachary, 1977). O banco de dados consiste em 34 indivíduos de um clube de karatê e as e as relações sociais entre os praticantes do esporte. Durante o período do estudo, ocorreu um conflito entre o administrador (John A) e o instrutor da equipe (Sr Hi). Aqueles que tomaram o partido do Sr Hi se mantiveram no clube e formaram uma aliança. Por outro lado, os que eram contra o Sr Hi se distanciaram ou formaram outro clube karatê. Nosso objetivo é então ilustrar como o método apresentado pode ser aplicado na detecção destes dois grupos.

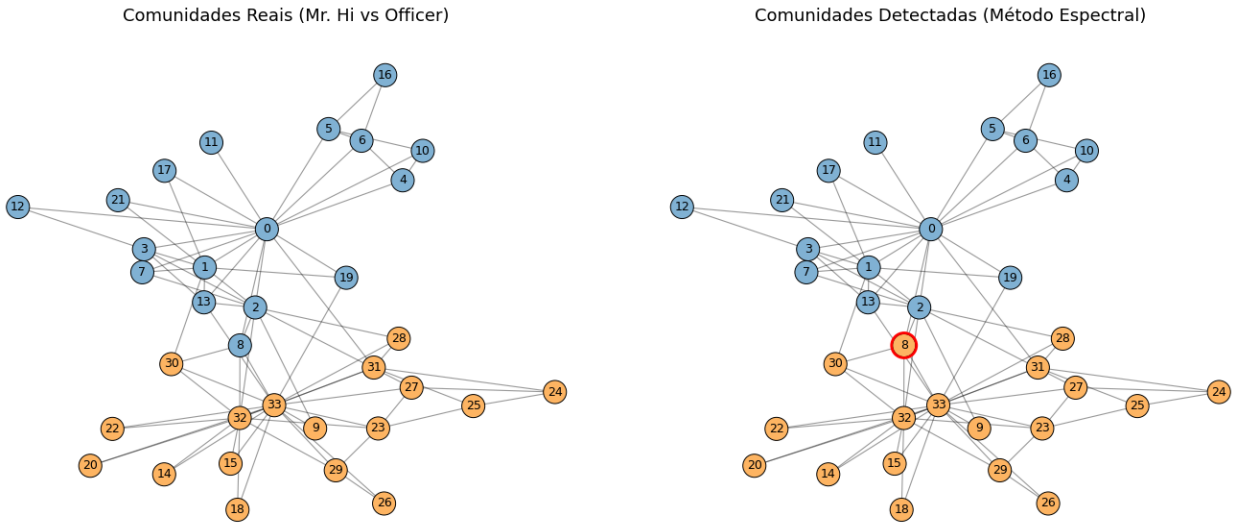

Figura 4.2: Comunidades reais à esquerda e comunidades detectadas pelo método espectral à direita.

O resultado pode ser observado na Figura 4.2. Os nós marcados com a cor azul se referem aos indivíduos que se mantiveram aliados ao Sr. Hi, enquanto os nós amarelos

representam a comunidade contra o Sr. Hi. Podemos observar que, comparando o grafo real e suas comunidades com as comunidades identificadas pelo algoritmo, apenas o nó 8 foi caracterizado incorretamente.

Concluímos então que, o estudo de desigualdade de concentração, no campo das matrizes aleatórias, é útil na fundamentação teórica de um método espectral. Como vimos, o modelo de blocos estocásticos, quando aliado ao algoritmo de detecção espectral apresentado, também pode ser eficiente no estudo de caso reais de detecção de duas comunidades.

Capítulo 5

Conclusão

Em síntese, este trabalho iniciou mostrou como desigualdades de concentração fornecem uma ponte rigorosa entre resultados clássicos da teoria da probabilidade e problemas centrais da estatística moderna e do aprendizado de máquina. A partir de desigualdades elementares, como as de Markov e Chebyshev, avançamos para resultados exponenciais mais fortes, como as desigualdades de Hoeffding e Chernoff, evidenciando como hipóteses adicionais sobre as variáveis aleatórias levam a garantias significativamente mais informativas. Em particular, vimos que essas ferramentas permitem quantificar de forma explícita o efeito do tamanho amostral e da complexidade do conjunto de hipóteses sobre o erro de generalização, oferecendo uma explicação matemática precisa para o fenômeno do *overfitting*. Dessa forma, o capítulo reforça o papel fundamental das desigualdades de concentração como instrumentos teóricos para compreender métodos de aprendizado estatístico.

Em seguida, investigamos o comportamento geométrico e probabilístico de vetores aleatórios em alta dimensão, destacando como a maldição da dimensão se manifesta por meio da concentração da norma, da quase ortogonalidade entre vetores aleatórios e da degradação de métodos baseados em distância. A partir do estudo de variáveis subgaussianas e subexponenciais, estabelecemos desigualdades de concentração fundamentais, como as de Hoeffding e Bernstein, que permitiram quantificar rigorosamente esses fenômenos de forma não assintótica. Em particular, mostramos que vetores com coordenadas subgaussianas se concentram em uma casca fina da esfera de raio proporcional a $\sqrt{n}$ e que produtos internos entre vetores isotrópicos independentes tendem a zero quando normalizados, evidenciando a geometria peculiar dos espaços de alta dimensão. Por fim, o Lema de Johnson–Lindenstrauss forneceu uma resposta construtiva a esses desafios, demons-

trando que projeções aleatórias permitem reduzir drasticamente a dimensionalidade dos dados preservando, com alta probabilidade, a estrutura métrica essencial. Esses resultados estabelecem a base conceitual e técnica para o estudo de matrizes aleatórias, tema do último capítulo.

Por fim, estudamos que resultados de concentração para matrizes aleatórias fornecem uma base teórica sólida para métodos espectrais em problemas de inferência estrutural em redes. A partir da decomposição em valores singulares e da interpretação da norma do operador como medida de perturbação espectral, foi possível quantificar rigorosamente o efeito do ruído aleatório por meio de desigualdades de concentração para matrizes subgaussianas. Em particular, a desigualdade de Davis–Kahan desempenhou um papel central ao permitir controlar a estabilidade dos autovetores sob perturbações, conectando a separação espectral da matriz determinística ao desempenho do estimador espectral construído a partir da matriz observada. Essa análise mostrou que, no modelo de blocos estocásticos, a estrutura de comunidades pode ser recuperada com alta probabilidade quando o sinal informativo domina o ruído aleatório, isto é, quando há separação suficiente entre os autovalores relevantes. A aplicação ao conjunto de dados do *Zachary's Karate Club* ilustrou, de forma concreta, como esses resultados teóricos se traduzem em um algoritmo eficiente de detecção de comunidades, reforçando a utilidade prática da teoria de matrizes aleatórias no estudo de redes complexas.

Referências Bibliográficas

Axler, S. (2015). *Linear Algebra Done Right*. Springer, third edition.

Bach, F. (2024). *Learning theory from first principles*. MIT press.

Bellman, R. (1966). Dynamic programming. *Science*, **153**(3731), 34–37.

Bollobás, B. (1998). *Modern graph theory*, volume 184. Springer Science & Business Media.

Boucheron, S., Lugosi, G. e Bousquet, O. (2003). Concentration inequalities. Em *Summer school on machine learning*, páginas 208–240. Springer.

Devroye, L. e Lugosi, G. (2001). Combinatorial methods in density estimation. Em *Springer Series in Statistics*.

Hastie, T., Tibshirani, R. e Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, second edition. ISBN 978-0-387-84857-0.

Horn, R. A. e Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press, second edition.

Mohri, M., Rostamizadeh, A. e Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Shalev-Shwartz, S. e Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA. ISBN 1107057132.

Tao, T. (2012). Topics in random matrix theory.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wainwright, M. J. (2019a). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wainwright, M. J. (2019b). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, **33**(4), 452–473.

Apêndice A

Desigualdades e identidades importantes

Resultado A.1 (Fórmulas da integral da cauda)**.** *Para toda variável aleatória X não negativa vale que*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t)\, dt. \tag{A.1}$$

De forma mais geral, para toda variável aleatória X (não necessariamente não negativa) e todo $p > 0$, vale que

$$\mathbb{E}[|X|^p] = \int_0^\infty \mathbb{P}(|X| > t)\, pt^{p-1} dt. \tag{A.2}$$

Demonstração. Qualquer número x não negativo pode ser expresso por

$$x = \int_0^x 1 dt = \int_0^\infty \mathbb{1}(t < x) dt.$$

Substituindo a variável aleatória X pelo número na expressão acima podemos, usando o teorema de Fubini-Tonelli, tomar a esperança dos dois lados:

$$\mathbb{E}[X] = \mathbb{E}\int_0^\infty \mathbb{1}(X > t) dt = \int_0^\infty \mathbb{E}[\mathbb{1}(X > t)]\, dt = \int_0^\infty \mathbb{P}(X > t)\, dt,$$

isto conclui a demonstração de A.1. Para provar A.2 podemos simplesmente fazer uma substituição de variáveis, considerando $|X|$ como a variável X na expressão acima e to-

mando $t = u^p$, obtendo

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty \mathbb{P}\left(|X|^p > t\right) dt && \text{(Afirmação A.1)} \\
&= \int_0^\infty \mathbb{P}\left(|X|^p > u^p\right) dt && (t = u^p) \\
&= \int_0^\infty \mathbb{P}\left(|X| > u\right) pu^{p-1} du.
\end{aligned}$$

□

Resultado A.2 (Fórmula de Stirling)**.** *Para todo n inteiro positivo vale que*

$$\left(\frac{n}{e}\right)^n \leq n! \leq en\left(\frac{n}{e}\right)^n.$$

Demonstração. Expandindo e^x em uma série de Taylor e ignorando todos os termos exceto o n-ésimo, temos que $e^x \geq x^n/n!$. Tomando $x = n$ e rearranjando os termos obtemos o limitante inferior. Para o limitante superior, note que

$$\log n! = \sum_{k=1}^{n} \log k \leq \int_1^n \log x dx + \log n = n(\log n - 1) + 1 + \log n.$$

A desigualdade acima vale quando comparamos as áreas geradas pela soma e pela integral como no teste da integral usual. Tomando a exponencial dos dois lados temos o limitante superior da Fórmula de Stirling. □

Resultado A.3. *Para todo $x \in \mathbb{R}$ vale que $1 + x \leq e^x$.*

Demonstração. Para provar a desigualdade usaremos um método elementar de cálculo. Definindo a função $f : \mathbb{R} \to \mathbb{R}$ por

$$f(x) = e^x - x - 1,$$

queremos provar que $f(x) \geq 0$ para todo x. Para isso, primeiro calculamos as derivadas $f'(x) = e^x - 1$ e $f''(x) = e^x > 0$. Resolvendo $f'(x) = 0$, obtemos que $x_c = 0$ é um ponto de mínimo de f, logo $f(0) = 0$ é o menor valor assumido pela função, exatamente como queríamos provar. □

Uma aplicação interessante deste lema consiste em provar facilmente que a série $\sum_{k=1}^{\infty} 1/k$ diverge. De acordo com o lema, para todo inteiro positivo n, temos $e^{1/n} \geq 1 + 1/n = (n+1)/n$, então

$$e^{1+1/2+\ldots+1/n} \geq \left(\frac{2}{1}\right)\left(\frac{3}{2}\right)\cdots\left(\frac{n}{n-1}\right)\left(\frac{n+1}{n}\right) = n+1.$$

Portanto $\sum_{k=1}^{n} 1/k \geq \log(n+1)$, isto é, a sequência de somas parciais diverge quando $n \to \infty$.

Resultado A.4. *Para todo $x \in [0, 1/2]$, vale que $1/(1-x) \leq e^{2x}$.*

Demonstração. Defina a função $f : \mathbb{R} - \{1\} \to \mathbb{R}$ por

$$f(x) = e^{2x} - \frac{1}{1-x}$$

e note que $f(0) = 0$. Iremos mostrar que, no intervalo $[0, 1/2]$, f é não decrescente. Começando com a derivada

$$f'(x) = 2e^{2x} - \frac{1}{(1-x)^2},$$

desejamos deduzir que, neste intervalo, $f'(x) \geq 0$. Para isso, escrevemos $f'(x)$ em uma diferença de quadrados da seguinte forma:

$$\begin{aligned} 2e^{2x} - \frac{1}{(1-x)^2} \geq 0 &\iff 2e^{2x}(1-x)^2 \geq 1 \\ &\iff \left(\sqrt{2}e^x(1-x)\right)^2 - 1^2 \geq 0 \\ &\iff \left(\sqrt{2}e^x(1-x)+1\right)\left(\sqrt{2}e^x(1-x)-1\right) \geq 0. \end{aligned}$$

Portanto, para provar que $f'(x) \geq 0$, basta deduzir que a última expressão no desenvolvimento é válida. O termo $\sqrt{2}e^x(1-x)+1$ é claramente positivo pois $x < 1$. O termo $\sqrt{2}e^x(1-x) - 1$ também é positivo no intervalo considerado, pois $e^x(1-x) > 1/\sqrt{2}$. Para provar esse fato, observe que $(e^x(1-x))' = -xe^x$, ou seja, $e^x(1-x)$ é uma função decrescente quando $x \geq 0$. Sendo assim seu menor valor em $[0, 1/2]$ será

$$e^{1/2}(1-1/2) = \frac{\sqrt{e}}{2} > \frac{\sqrt{2}}{2} = \frac{1}{\sqrt{2}},$$

isto conclui a demonstração da afirmação de que a derivada de f é não negativa e, portanto, do lema. □

Resultado A.5. *Para todo $x \in \mathbb{R}$, vale que $e^x \leq x + e^{x^2}$.*

Demonstração. Defina

$$f(x) = x + e^{x^2} - e^x.$$

Mostraremos $f(x) \geq 0$ dividindo em três casos.

1. **Caso $x \geq 1$.**
   Como $x \geq 1$ implica $x^2 \geq x$ e a exponencial é crescente, segue

   $$e^{x^2} \geq e^x.$$

   Assim

   $$f(x) = x + e^{x^2} - e^x \geq x \geq 0.$$

2. **Caso $0 \leq x \leq 1$.**
   Considere $\varphi(t) = 1+t+t^2-e^t$. Tem-se $\varphi(0) = 0$ e $\varphi(1) = 3-e > 0$. Uma verificação curta das derivadas (ou desenvolvimento de Taylor com controle do resto) mostra que $\varphi(t) \geq 0$ para todo $t \in [0, 1]$, ou seja

   $$e^t \leq 1 + t + t^2 \quad \text{para } t \in [0, 1].$$

   Aplicando com $t = x$ e usando $e^{x^2} \geq 1 + x^2$ (pois $e^u \geq 1 + u$ para todo u),

   $$x + e^{x^2} \geq x + (1 + x^2) = 1 + x + x^2 \geq e^x,$$

   logo $f(x) \geq 0$.

3. **Caso $x < 0$.**
   Pela convexidade da exponencial vale $e^t \geq 1 + t$ para todo $t \in \mathbb{R}$. Em particular $e^x \geq 1+x$, de modo que $e^x - x - 1 \geq 0$. Como $x < 0$ temos $x^2 > -x$, logo $e^{x^2} > e^{-x}$; além disso $e^x \leq e^{-x}$. Portanto

   $$e^{x^2} - 1 > e^{-x} - 1 \geq e^x - x - 1,$$

   o que implica

   $$e^{x^2} - e^x + x \geq 0,$$

isto é, $f(x) \geq 0$.

□

Resultado A.6. *Para todo $x \in \mathbb{R}$ e $p > 0$, vale que*

$$|x|^p \leq p^p(e^x + e^{-x}).$$

Demonstração. Note que

$$e^x + e^{-x} \geq e^x \Rightarrow (e^x + e^{-x})^{1/p} \geq e^{x/p} \geq x/p + 1 \geq x/p.$$

□