Entropia: introdução à Teoria Matemática da (des)Informação [1]

Bernardo N. B. Lima,
Leandro Martins Cioletti,
Marcelo de O. Terra Cunha,
Gastão A. Braga

Departamento de Matemática - UFMG
CP 702 - Belo Horizonte - 30123-970

15 de outubro de 2004

[1]Minicurso a ser apresentado na II Bienal da SBM - UFBa - Salvador - 25 a 29/10/2004

Introdução

Entropia

O conceito central dessas notas nasceu em meados do século XIX, quando a *Termodinâmica* se desenvolvia. A Termodinâmica foi a teoria científica destinada a explicar e desenvolver as máquinas térmicas, ponto central da Revolução Industrial. Nesse contexto, a *entropia* aparece com o intuito de distinguir processos *reversíveis* e *irreversíveis*, estando assim intimamente ligada com a questão da direção da chamada "seta do tempo". A *Segunda Lei da Termodinâmica* aponta exatamente para o crescimento da entropia de um sistema. A definição de entropia a essa altura era como uma relação entre diferenciais, onde o inverso da temperatura jogava o papel de *fator integrante* para a quantidade de calor. É raro encontrar alguém que se sinta realmente satisfeito com tal definição.

A primeira grande reinterpretação do conceito veio com a Mecânica Estatística, teoria que busca conciliar duas outras aparentemente incompatíveis: a Termodinâmica, com sua Segunda Lei, e a Mecânica, onde se aplica a reversibilidade temporal. O conceito de entropia em mecânica estatística está ligado ao logaritmo da contagem de quantos estados microscopicamente distintos são compatíveis com uma mesma descrição macroscópica do sistema. Assim se dá a conciliação entre Mecânica e Termodinâmica: os processos envolvendo diminuição de entropia não são impossíveis, mas "apenas" a probabilidade que eles ocorram é extremamente pequena. Quando lembramos do número de Avogrado (6×10^{23}) temos uma noção melhor do que queremos dizer com "extremamente pequena"). É uma curiosidade histórica que a expressão resumindo isso,

$$S = k \ln W,$$

está na lápide do túmulo de Ludwig Boltzmann (1844-1906), seu criador.

Um quarto de século depois, a *Mecânica Quântica* toma forma, e John von Neumann (1905-1957) faz a ligação do conceito de entropia com esta nova teoria de fenômenos

microscópicos. Para a ciência e a tecnologia a mecânica quântica foi uma revolução (transistor, laser e energia nuclear são exemplos de aplicações), mas até então, para o conceito de entropia foi apenas uma generalização natural. Hoje, porém, podemos considerar esta contribuição de von Neumann como uma preparação para a posterior teoria quântica da informação.

Outra revolução no conceito surge no pós-guerra. Com as telecomunicações ganhando força, faltava arcabouço teórico capaz de predizer a capacidade de um canal de comunicação. De uma forma simplificada, como determinar a bitola mínima de um fio que unisse duas centrais telefônicas de maneira eficiente, ou, visto por outro lado, como melhor aproveitar os recursos disponíveis. Em 1948, Claude E. Shannon (1916-2001) apresentou esta teoria, e o conceito central foi a *entropia de uma fonte de informação*. Nascia aí a *Teoria da Informação*, e o conceito de entropia ganhava uma nova faceta: agora ele dizia como armazenar e transmitir informação de maneira mais econômica. Com ele ficou claro que a noção de entropia cabia bem no contexto de *probabilidades*, e não necessariamente em teorias físicas como Termodinâmica ou Mecânica Estatísitica (clássica ou quântica). De certa forma, sua presença era assegurada pelos métodos estatísticos e não pelos conceitos mecânicos da teoria.

Devemos citar também que na última metade do século XX ganharam importância a noção de taxa de criação de entropia em *sistemas dinâmicos*, bem como surgiu o conceito de *entropia não-extensiva*, devido a Constantino Tsallis (1943 -). A primeira busca uma compreensão mais detalhada de como sistemas dinâmicos caminham em direção ao equilíbrio, enquanto o segundo se mostra adequado para contextos onde as variáveis envolvidas não são independentes.

A mais recente reformulação de conceitos vem nas últimas décadas do século XX, e ainda se expande: o surgimento da *Teoria Quântica da Informação*. Duas visões complementares levam a esta teoria: por um lado a idéia de buscar vantagens em aplicar mecânica quântica ao tratamento da informação, por outro a sua necessidade devido ao processo de miniaturização dos componentes eletrônicos, expresso na chamada Lei de Moore. No primeiro caso, há o exemplo do Algoritmo de Shor, que faz a fatoração de inteiros em tempo polinomial, o que tornaria "fácil" quebrar a criptografia RSA. Usamos "tornaria" pois não há *computador quântico* operando com grande quantidade de bits quânticos, como seria exigido para uma aplicação como esta. Uma outra conquista da teoria quântica da informação é a *criptografia quântica*, processos de distribuição pública de chaves privadas, baseados na mecânica quântica, e não na provável complexidade de certos algoritmos, como o de fatoração. A criptografia quântica já é realidade, em vias de se tornar comercial[10]. A demonstração de sua segurança envolve o conceito de entropia, que será aqui abordado.

O texto

No segundo semestre de 2003 os autores deste texto, junto com Ricardo Falcão e Daniel Cavalcanti, realizaram seminários no intuito de buscar linguagem comum e conhecimento em Teoria Quântica da Informação. O texto base para isso foi [5]. No início de 2004, contando ainda com a participação de Paulo Cupertino de Lima, chegamos ao conceito de entropia, onde ficava claro que a bagagem de cada um trazia visões complementares sobre o assunto. Por outro lado, não é comum encontrar essa bagagem reunida. Entropia aparece como conceito nos livros das diferentes teorias, mas só em alguns poucos textos é o fio condutor. Surgiu assim a idéia de preparar um texto como o aqui apresentado, imediatamente pensado para a apresentação de um mini-curso na II Bienal de Matemática, exigindo do público alvo apenas o conhecimento comum ao ciclo básico de ciências exatas: cálculo de funções de várias variáveis e álgebra linear.

Como o pano de fundo é a noção de *probabilidade*, o primeiro capítulo que se segue apresenta essa teoria na sua visão moderna, sem a necessidade de pré-requisitos sofisticados. O preço que se paga é restringir a teoria a *espaços amostrais finitos*, mas houve a preocupação ao longo do texto de, sempre que possível, manter a discusão em termos mais gerais, embora as definições e exemplos diretamente trabalhados obedeçam tal restrição.

O capítulo seguinte traz a *teoria clássica da informação*, apresentando algumas propriedades importantes das diversas entropias que seguem da primeira definição (noções como entropia conjunta, entropia relativa e informação mútua), e culminando com o Teorema de Shannon, onde surge a relação entre entropia de uma distribução e a possibilidade de compactar dados apresentados segundo esta distribuição.

O capítulo 3 trata do conceito de entropia em *teoria quântica da informação*. É natural, no espírito deste texto, fazer uma introdução dos ingredientes necessários para o tratamento do assunto, assim os conceitos básicos de mecânica quântica são apresentados, assumindo apenas que o leitor conhece álgebra linear de espaços vetoriais com dimensão finita. É claro que não cabe aqui a discussão de detalhes da teoria, mas alguns dos seus pontos mais curiosos são tocados ao longo do texto. O capítulo traz as generalizações naturais dos conceitos clássicos, mas aponta também algumas das importantes diferenças entre a teoria clássica e a quântica. O capítulo se encerra com o Teorema de Schumacher, a perfeita tradução do Teorema de Shannon para este contexto.

O capítulo 4 trata da *mecânica estatística de equilíbrio* em redes finitas. Este é o cenário onde se torna mais elementar obter resultados rigorosos, cuja extensão a redes infinitas e a modelos mais gerais é ainda, em grande parte, objeto de pesquisa atual. Neste capítulo

demonstramos que para cada valor de uma certa *energia livre*, o estado de equilíbrio é aquele que maximiza a entropia sujeito a tal restrição.

O texto é auto-contido, e assume apenas os pré-requisitos já comentados, além de disposição. Os exercícios são considerados parte indispensável do texto. Vários desejos foram abandonados em nome da premência de tempo: não incluímos um apêndice sobre criptografia quântica, não escrevemos um capítulo sobre entropia em sistemas dinâmicos e não conseguimos passar o texto antes para um número considerável de colegas que pudessem nos ajudar a ter um texto mais definitivo. Nesse sentido, vemos estas notas como uma primeira versão de algo que pode se tornar mais completo. Ainda assim, achamos que servem bem como convite ao estudo de um conceito naturalmente interdisciplinar, e esperamos que o leitor tenha prazer nas páginas que se seguem.

Sumário

Introdução **iii**

Entropia iii

O texto v

1 Noções de Probabilidade **1**

1.1 Espaços de Probabilidade 1

1.2 Probabilidade Condicional e Independência 3

1.3 Variáveis Aleatórias 5

1.4 Esperança e Lei dos Grandes Números 8

2 Entropia e Teoria Clássica da Informação **13**

2.1 Definições e primeiros exemplos 13

2.2 Concavidade de $H(X)$ 15

2.3 Entropia Relativa 16

2.4 A Entropia Conjunta 18

2.5 Entropia Condicional e Informação Mútua 19

2.6 Compactação de Dados e o Teorema de Shannon 20

3 Entropia e Teoria Quântica da Informação 25

3.1 Noções rápidas de Mecânica Quântica 25

3.1.1 Notação de Dirac 29

3.2 Entropia de von Neumann 30

3.3 Entropia e Medições 32

3.4 Sistemas quânticos compostos 34

3.5 Entropia em sistemas compostos 37

3.6 Entropia de uma Mistura de Estados Quânticos 40

3.7 Teorema de Schumacher 41

4 Entropia e Mecânica Estatística 47

4.1 Definições Preliminares 48

4.2 O Princípio Variacional 50

Referências Bibliográficas 55

Índice Remissivo 56

Capítulo 1

Noções Básicas de Probabilidade em Espaço Amostral Finito

O objetivo deste capítulo é introduzir os resultados e definições básicos da Teoria da Probabilidade que serão necessários para a compreensão do restante do texto. Iremos nos restrigir ao caso (bem mais simples!) onde o espaço amostral é um conjunto finito. Para os interessados, algumas referências de textos em probabilidade são [1] (esta é clássica e existe tradução para o Português), [2] e [6] (cujo ótimo primeiro capítulo é um curso completo de probabilidade em espaços amostrais finitos, onde tópicos sofisticados da Teoria da Probabilidade são abordados, como o Teorema Central do Limite, a Lei dos Grandes Números, Cadeias de Markov e Martingais).

1.1 Espaços de Probabilidade

Atualmente, as definições de Probabilidade e de Espaço de Probabilidade empregadas na comunidade científica são definições axiomáticas: são o que conhecemos hoje como os "Axiomas de Kolmogorov". Antes de descrevê-los, será importante definir o conceito de σ-álgebra de um conjunto, a saber:

Definição 1.1 *Seja Ω um conjunto qualquer e $\mathcal{A}$ uma classe de subconjuntos de Ω, dizemos que $\mathcal{A}$ é uma* σ-álgebra *de Ω, se $\mathcal{A}$ satisfaz às seguintes condições:*
i) $\emptyset \in \mathcal{A}$;

ii) Se $A \in \mathcal{A}$*, então* $A^c \in \mathcal{A}$*;*
iii) Se $A_1, \ldots, A_n, \ldots \in \mathcal{A}$*, então* $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$*.*

Definição 1.2 (Axiomas de Kolmogorov) *Um* Espaço de Probabilidades *é uma tripla ordenada* $(\Omega, \mathcal{A}, P)$*, onde* Ω *é um conjunto qualquer não-vazio, denominado* Espaço Amostral*,* $\mathcal{A}$ *é uma* σ*-álgebra de subconjuntos de* Ω *e* P *é uma função* $P : \mathcal{A} \to \mathbb{R}$ *satisfazendo:*
i) $P(A) \geq 0, \ \forall \ A \in \mathcal{A}$*;*
ii) $P(\Omega) = 1$*;*
iii) Se $A_1, \ldots, A_n, \ldots \in \mathcal{A}$ *e são dois a dois disjuntos , então* $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$*.*

Neste caso, dizemos que a função P *é uma* medida de probabilidade*,(ou simplesmente* probabilidade*) e os elementos de* $\mathcal{A}$*, o domínio de* P*, serão chamados de* eventos*.*

Durante todo o texto, iremos apenas considerar Espaços de Probabilidade em que o Espaço Amostral seja um conjunto finito (para uma definição precisa de conjunto finito, veja o capítulo 2 de [4]). No caso Ω finito, podemos sempre tomar a σ-álgebra $\mathcal{A}$ como sendo $\mathbb{P}(\Omega)$, o conjunto das partes de Ω (i.e., aquele formado por todos os subconjuntos de Ω).

Exercício 1.1 *Denotando por* $\#\Omega$*, a cardinalidade do conjunto* Ω*, mostre que se* $\#\Omega = n$*, então* $\#\mathbb{P}(\Omega) = 2^n$*.*

Se $\Omega = \{\omega_1, \ldots, \omega_n\}$, então associamos a toda medida de probabilidade P, de modo biunívoco, um *vetor de probabilidade*, ou *vetor de distribuição de probabilidade*, $(p_1, \ldots, p_n)$, onde $p_i \geq 0, \ i = 1, \ldots n$, e $p_1 + \cdots + p_n = 1$. Neste caso, temos que

$$P(A) = \sum_{i;\omega_i \in A} p_i, \ \forall \ A \in \mathbb{P}(\Omega) \tag{1.1}$$

e p_i é a probabilidade de ocorrência do evento $\{\omega_i\}$. Em termos menos formais, se formos realizar um sorteio onde Ω seja um conjunto que contenha todas as possíveis realizações do nosso sorteio, considerar uma distribuição de probabilidade dada pelo vetor $(p_1, \ldots, p_n)$ equivale a assumir que p_i é a probabilidade do resultado do sorteio ser ω_i.

Exercício 1.2 *Verifique que* P*, definida por (1.1) é de fato uma medida de probabilidade!*

Um caso particular de extrema importância é quando assumimos que todos os possíveis resultados do nosso sorteio são equiprováveis, isto é, $p_i = \frac{1}{n},\ i = 1, \ldots n$. Neste caso, temos que a probabilidade de qualquer evento A pertencente a $\mathbb{P}(\Omega)$ é dada por

$$P(A) = \frac{\#A}{\#\Omega},$$

ou seja, o problema do cálculo de uma probabilidade se resume a um problema de contagem, já que a probabilidade de ocorrência de qualquer evento (conjunto) A é dada pela proporção entre o número de elementos de A e o número de elementos do Espaço Amostral.

Exercício 1.3 *i) n pessoas participam de um "amigo oculto", isto é, existe uma urna contendo n bolas com os nomes de cada um dos n participantes e cada um dos participantes sorteia, sem repor, uma bola com o nome do seu "amigo oculto"; deste modo, cada participante terá um único "amigo oculto"e será "amigo oculto"de uma única pessoa. Calcule a probabilidade de haver pelo menos uma pessoa que sorteou a si próprio como "amigo oculto".*
ii) Mostre que o limite da resposta do ítem anterior quando $n \to +\infty$ *é* $1 - e^{-1}$.

1.2 Probabilidade Condicional e Independência

Sejam $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade, A e B eventos tais que $P(B) \neq 0$. Definimos a *probabilidade condicional do evento A, dado o evento B*, como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Intuitivamente, podemos pensar em $P(A|B)$ como a proporção que o evento $A \cap B$ ocupa no evento B. A seguir enunciaremos os principais resultados sobre probabilidades condicionais:

Teorema 1.1 (Teorema da Multiplicação) *Sejam* $A_1, A_2, \ldots, A_n$ *eventos quaisquer. Então:*

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1) \times P(A_2|A_1) \times \cdots \times P(A_n|A_1 \cap A_{n-1}).$$

Demonstração: O caso $n = 2$ segue da definição de probabilidade condicional. Para os demais valores de n, a prova segue facilmente por indução finita. □

Definição 1.3 *Seja $\mathcal{P} = \{A_1, \ldots, A_n\}$ um conjunto de eventos. Dizemos que o conjunto $\mathcal{P}$ é uma* partição *do espaço amostral Ω se:*
i) $A_i \cap A_j = \emptyset$, $\forall\ i \neq j$, isto é, os elementos de $\mathcal{P}$ são conjuntos dois a dois disjuntos;
ii) $\cup_{i=1}^{n} A_i = \Omega$.

Seja $\mathcal{P} = \{A_1, \ldots, A_n\}$ uma partição do espaço amostral Ω e B um evento qualquer. Então temos os seguintes resultados:

Teorema 1.2 (Teorema da Probabilidade Total)

$$P(B) = \sum_{i=1}^{n} P(A_i) \times P(B|A_i).$$

Demonstração: Como $\mathcal{P}$ é uma partição do espaço amostral Ω, podemos escrever B como a seguinte união disjunta

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_n \cap B)$$

Portanto,

$$P(B) = \sum_{i=1}^{n} P(A_i \cap B) = \sum_{i=1}^{n} P(A_i) \times P(B|A_i).$$

□

Corolário 1.1 (Fórmula de Bayes)

$$P(A_i|B) = \frac{P(A_i) \times P(B|A_i)}{\sum_{j=1}^{n} P(A_j) \times P(B|A_j)},\ \forall\ i = 1, \ldots, n.$$

Definição 1.4 *Sejam A e B dois eventos quaisquer em um mesmo espaço de probabilidades. Dizemos que A e B são independentes se*

$$P(A \cap B) = P(A) \times P(B).$$

Se A e B são independentes com $P(A) \neq 0$ e $P(B) \neq 0$ podemos ver facilmente que $P(B|A) = P(B)$ e $P(A|B) = P(A)$, isto é, a ocorrência do evento A não afeta a ocorrência do evento B e vice-versa.

Exercício 1.4 *i) Sejam A e B eventos independentes. Mostre que os pares de eventos A e B^c, A^c e B e A^c e B^c são pares de eventos independentes.*
ii) Mostre que o evento A é independente de si mesmo se, e somente se, $P(A) = 0$ ou 1.

Definição 1.5 *Sejam $A_1, A_2, \ldots, A_n$ eventos quaisquer em um mesmo espaço de probabilidades. Dizemos que $A_1, A_2, \ldots, A_n$ são dois a dois independentes se*

$$P(A_i \cap A_j) = P(A_i) \times P(A_j), \ \forall \ i \neq j.$$

Dizemos que $A_1, A_2, \ldots, A_n$ são coletivamente independentes se para todo $k = 2, \ldots, n$,

$$P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \times \cdots \times P(A_{i_k}), \ \forall \ 1 \leq i_1 < \cdots < i_k \leq n.$$

Portanto, toda família de eventos coletivamente independente é dois a dois independente; porém a recíproca é falsa conforme podemos ver no exemplo abaixo.

Exemplo 1.1 *Considere o experimento aleatório de lançar duas vezes uma mesma moeda honesta. Sejam A, B e C os eventos cara no primeiro lançamento, cara no segundo lançamento e o resultado dos dois lançamentos é o mesmo, respectivamente. Podemos observar que os eventos A, B e C são dois a dois independentes mas não são coletivamente independentes, pois*

$$P(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = P(A) \times P(B) \times P(C).$$

1.3 Variáveis Aleatórias

Seja $(\Omega, \mathcal{A}, P)$ um espaço de probabilidade. A função $X : \Omega \to \mathbb{R}$ é uma variável aleatória se

$$\{\omega \in \Omega; X(\omega) \leq x\} \in \mathcal{A}, \ \forall \ x \in \mathbb{R}.$$

Observe que no caso de espaço amostral finito, como tomamos $\mathcal{A} = \mathbb{P}(\Omega)$, temos que toda função $X : \Omega \to \mathbb{R}$ é uma variável aleatória. Como estamos supondo sempre o caso $\#\Omega < +\infty$, será útil considerarmos funções $X : \Omega \to \Lambda$ onde Λ é um conjunto finito de símbolos que não é necessariamente subconjunto de $\mathbb{R}$; mesmo assim chamaremos tais funções de variáveis aleatórias.

Definição 1.6 *Seja $X : \Omega \to \Lambda$ uma variável aleatória. A* distribuição de probabilidades em $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$ induzida por X*, ou simplesmente, a* distribuição de X *é o vetor de probabilidade em Λ, $P_X = (x_1, \ldots, x_m)$ onde*

$$x_i = P(X = \lambda_i) = \sum_{j; X(\omega_j) = \lambda_i} p_j, \ \forall \ i = 1, \ldots, m.$$

Exemplo 1.2 *Seja $X : \Omega \to \Lambda$ variável aleatória onde $\#\Lambda = 2$. Neste caso dizemos que X é uma* variável aleatória de Bernoulli. *Podemos pensar que os valores assumidos por X podem ser "sucesso"e "fracasso", ou 0 e 1, ou cara e coroa etc. dependendo do contexto. A distribuição de X é determinada por um único parâmetro, p, que é a probabilidade de X ser "sucesso", logo a probabilidade de X ser "fracasso"é $1 - p$.*

Exemplo 1.3 *Seja $A \subset \Omega$. A função $I_A : \Omega \to \{0, 1\}$, onde*

$$I_A(\omega) = \begin{cases} 1, & \text{se } \omega \in A \\ 0, & \text{se } \omega \notin A. \end{cases} \tag{1.2}$$

é uma variável aleatória e I_A é chamada de função indicadora do conjunto A.

Exercício 1.5 *Mostre que se $\#\Omega < \infty$, toda variável aleatória $X : \Omega \to \mathbb{R}$ pode ser escrita na forma*

$$X = \sum_{i=1}^{n} x_i I_{A_i},$$

onde $X(\Omega) = \{x_1, \ldots, x_n\}$ e $A_i = \{\omega \in \Omega : X(\omega) = x_i\}$.

Definição 1.7 *Seja $X : \Omega \to \Lambda$ variável aleatória onde $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$. Dizemos que X tem* distribuição uniforme*, ou* equiprovável*, em Λ se a distribuição de X é o vetor $P_X = (\frac{1}{m}, \ldots, \frac{1}{m})$.*

O conceito de independência pode ser estendido para variáveis aleatórias conforme a definição a seguir:

Definição 1.8 *Sejam $X : \Omega \to \Lambda$ e $Y : \Omega \to \Xi$ variáveis aleatórias. Dizemos que o par de variáveis aleatórias X e Y é independente se, $\forall\ A \subset \Lambda$ e $\forall\ B \subset \Xi$, os eventos $X^{-1}(A)$ e $Y^{-1}(B)$ são independentes. De modo análogo, podemos estender as definições de independência dois a dois e independência coletiva para conjuntos de variáveis aleatórias definidas em um mesmo espaço amostral.*

Exercício 1.6 *Sejam I_A e I_B as funções indicadoras dos eventos A e B. Mostre que as variáveis aleatórias I_A e I_B são independentes se, e somente se, os eventos A e B são independentes.*

Exemplo 1.4 *Sejam $X_1, \ldots, X_n$ variáveis aleatórias de Bernoulli (coletivamente) independentes e com a mesma distribuição. Considere a variável aleatória $S = X_1 + \cdots + X_n$, que mede o número de sucessos ocorridos durante a repetição de n ensaios independentes de Bernoulli. Dizemos que S possui* distribuição binomial com parâmetros n e p, *ou $S \sim b(n, p)$, onde p é a probabilidade de "sucesso"em cada ensaio de Bernoulli.*

Exercício 1.7 *Verifique que*

$$P(S = i) = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}, \ \forall \ i = 0, \ldots, n.$$

O exemplo anterior, e muitas outras situações importantes em Probabilidade, diz respeito a propriedades de seqüências de variáveis aleatórias independentes. Dentre os diversos tipos de dependência que podem ser introduzidos, definiremos abaixo o que é conhecido como Cadeias de Markov.

Definição 1.9 *i) Seja $(X_n)_{n \in \mathbb{N}}$, $X_n : \Omega \to \Lambda$ uma seqüência de variáveis aleatórias (é claro que $\#\Lambda < \infty$). Dizemos que tal seqüência forma uma* Cadeia de Markov *se*

$$P(X_{n+1} = x_{n+1} | X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} = x_{n+1} | X_n = x_n)$$

para todo $n \in \mathbb{N}$ e para toda seqüência $(x_i)_{i=1}^{n+1}$ tal que $x_i \in \Lambda$.
ii) Tal seqüência forma uma Cadeia de Markov homogênea *(no tempo), se*

$$P(X_{n+1} = y | X_n = x) = P(X_2 = y | X_1 = x), \ \forall \ n \in \mathbb{N}, \ \forall \ x, y \in \Lambda.$$

Em palavras, em uma Cadeia de Markov homogênea quaisquer que sejam os estados x e y, a probabilidade condicional, de uma variável aleatória da seqüência assumir no tempo futuro $(n+1)$ o estado y dada toda a seqüência de estados passados e presente (até o tempo n) assumidos pela seqüência de variáveis aleatórias, depende apenas dos estados presente e futuro, x e y, não dependendo da posição (temporal), n, e nem dos estados passados.

Exercício 1.8 *Mostre que a seqüência de variáveis aleatórias (X, Y, Z) forma uma Cadeia de Markov homogênea se, e somente se, a seqüência (Z, Y, X) também é uma Cadeia de Markov homogênea.*

1.4 Esperança e Lei dos Grandes Números

Definição 1.10 *Seja $X : \Omega \to \Lambda \subset \mathbb{R}$ uma variável aleatória, onde $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$. A* Esperança da variável aleatória *X, $E(X)$, é*

$$E(X) = \sum_{i=1}^{m} \lambda_i \cdot P(X = \lambda_i).$$

Exemplo 1.5 *Para os exemplos da seção anterior, pode-se verificar (exercício!) que*
i) Se $X = I_A$, a função indicadora do evento A, então $E(X) = P(A)$;
ii) Se $X : \Omega \to \Lambda$ tem distribuição uniforme, ou equiprovável, em Λ, então $E(X) = \frac{\lambda_1 + \cdots + \lambda_m}{m}$, a média aritmética dos valores de Λ;
iii) Se $X \sim b(n, p)$, então $E(X) = np$.

Teorema 1.3 *i) Sejam $X, Y : \Omega \to \mathbb{R}$ variáveis aleatórias. Então*

$$E(aX + bY) = aE(X) + bE(Y), \ \forall \ a, b \in \mathbb{R}.$$

ii) Se X e Y são independentes, então $E(XY) = E(X)E(Y)$.

Demonstração:*i*) Sejam $X = \sum_{i=1}^{n} x_i I_{A_i}$ e $Y = \sum_{j=1}^{m} y_j I_{B_j}$, onde $A_i = (X = x_i)$, $i = 1, \ldots, n$ e $B_j = (Y = y_j)$, $j = 1, \ldots, m$. Logo

$$E(aX + bY) = E(\sum_{i=1}^{n}\sum_{j=1}^{m}(ax_i + by_j)I_{A_i \cap B_j}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(ax_i + by_j)P(A_i \cap B_j) =$$

$$= \ a\sum_{i=1}^{n} x_i P(A_i) + b\sum_{j=1}^{m} y_j P(B_j) = aE(X) + bE(Y).$$

ii)

$$E(XY) = E(\sum_{i=1}^{n} x_i I_{A_i} \times \sum_{j=1}^{m} y_j I_{B_j}) = E(\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j I_{A_i \cap B_j}) =$$

$$= \ \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(A_i \cap B_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j P(A_i)P(B_j) = E(X)E(Y)$$

□

Exemplo 1.6 *Seja X a variável aleatória que assume os valores 0, $\frac{\pi}{2}$ e π com probabilidade $\frac{1}{3}$. Sejam Y e Z variáveis aleatórias definidas como $Y = \sin X$ e $Z = \cos X$. O leitor deve verificar que $E(YZ) = E(Y)E(Z)$; porém Y e Z não são independentes pois $P(Y = 1, Z = 1) \neq P(Y = 1)P(Z = 1)$. Logo, a recíproca da parte ii) do teorema anterior é falsa.*

Sejam X e Y variáveis aleatórias. Definimos a *covariância* entre X e Y como

$$cov(X, Y) = E(XY) - E(X)E(Y).$$

A *variância* da variável aleatória X é definida como

$$V(X) = cov(X, X) = E(X^2) - [E(X)]^2 = E[X - E(X)]^2.$$

Exercício 1.9 *Mostre que*
i)$V(X_1 + \cdots + X_n) = \sum_{i=1}^n V(X_i) + 2\sum_{1\leq i<j\leq n} cov(X_i, X_j)$. Em particular, se $(X_i)_{i=1}^n$ são dois a dois independentes $V(\sum_{i=1}^n X_i) = \sum_{i=1}^n V(X_i)$.
ii) Se $X \sim b(n, p)$, então $V(X) = np(1 - p)$.

Agora estamos prontos para enunciar (e provar!) um dos resultados mais importantes da Teoria da Probabilidade: a Lei dos Grandes Números. Em palavras, ela diz que sob certas hipóteses, ao realizarmos n repetições independentes de uma certa variável aleatória X, a média aritmética das n observações realizadas se aproxima de $E(X)$ quando o número de observações n cresce para o infinito.

Há diversas versões para a Lei dos Grandes Números. A que provaremos é uma versão da Lei Fraca dos Grandes Números que será suficiente para o que virá a seguir. Na demostração, a seguinte desigualdade será necessária.

Teorema 1.4 (Desigualdade de Chebyshev) *Seja $X : \Omega \to \mathbb{R}$ uma variável aleatória não negativa. Então*

$$P(X \geq \epsilon) \leq \frac{E(X)}{\epsilon}, \ \forall\ \epsilon > 0.$$

Demonstração: Novamente, seja X da forma

$$X = \sum_{i=1}^{n} x_i I_{A_i},$$

com $A_i = (X = x_i),\ i = 1, \ldots, n$. Então

$$E(X) = \sum_{i=1}^{n} x_i P(A_i) \geq \sum_{i;x_i \geq \epsilon} x_i P(A_i) \ \geq$$

$$\geq \ \epsilon \sum_{i;x_i \geq \epsilon} P(A_i) = \epsilon P(X \geq \epsilon).$$

□

Definição 1.11 *Dizemos que uma seqüência $(X_n)_{n\in\mathbb{N}}$ de variáveis aleatórias, têm variâncias uniformemente limitadas se, $\exists\ c > 0$ tal que $V(X_n) \leq c,\ \forall\ n \in \mathbb{N}$.*

Teorema 1.5 (Lei Fraca dos Grandes Números) *Sejam $(X_n)_{n\in\mathbb{N}}$ seqüência de variáveis aleatórias com variâncias uniformemente limitadas e com $cov(X_n, X_m) = 0$ se $n \neq m$. Então, para todo $\epsilon > 0$*

$$\lim_{n\to+\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \frac{E(X_1 + \cdots + X_n)}{n}\right| \geq \epsilon\right) = 0.$$

Demonstração:

$$P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \frac{E(X_1 + \cdots + X_n)}{n}\right| \geq \epsilon\right) \ =$$

$$= \ P\left(|(X_1 - E(X_1)) + \cdots + (X_n - E(X_n))|^2 \geq n^2\epsilon^2\right) \ \leq$$

$$\leq \ \frac{E\left(\sum_{i=1}^{n}(X_i - E(X_i))\right)^2}{n^2\epsilon^2} = \frac{\sum_{i=1}^{n} V(X_i) + 2\sum_{1\leq i<j\leq n} cov(X_i, X_j)}{n^2\epsilon^2} \leq \frac{c}{n\epsilon^2}.$$

Na primeira desigualdade acima, foi utilizada a desigualdade de Chebyshev e na segunda desigualdade, foram utilizadas as hipóteses assumidas pela seqüência $(X_n)_{n\in\mathbb{N}}$. Logo, segue a Lei Fraca dos Grandes Números. □

Corolário 1.2 *Seja $(X_n)_{n\in\mathbb{N}}$ seqüência de variáveis aleatórias independentes e identicamente distribuidas, com $V(X_1) < +\infty$. Então, para todo $\epsilon > 0$*

$$\lim_{n\to+\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - E(X)\right| \geq \epsilon\right) = 0.$$

Em palavras, a Lei dos Grandes Números diz que se realizarmos um grande número de observações independentes sobre a variável aleatória X, a média aritmética das observações converge para $E(X)$ quando o número de observações vai a infinito. Em particular, se X é a função indicadora I_A, podemos concluir que após n observações independentes do evento A, a freqüência relativa com que ele ocorre aproxima-se de $P(A)$. Esta afirmação, antes dos axiomas de Kolmogorov, era a base da *teoria freqüencista*, que dava à teoria de probabilidades um caráter mais empírico que matemático.

Capítulo 2

Entropia e Teoria Clássica da Informação

Este capítulo começa com a definição de nosso conceito central: *entropia*. Depois de explorar o caso de um único sistema, passamos a diversos conceitos naturais sobre sistemas compostos: *entropia conjunta*, *entropia relativa* e *informação mútua* são definidas e várias de suas propriedades são exploradas. O capítulo se encerra com o resultado central da teoria clássica da informação, o Teorema de Shannon, que relaciona entropia com compressão de dados.

2.1 Definições e primeiros exemplos

Definição 2.1 *Seja X uma variável aleatória com vetor de probabilidade $(p_1, \ldots, p_n)$. Definimos a* entropia de Shannon da variável aleatória X *como*

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i. \tag{2.1}$$

Como o leitor pode notar a entropia de Shannon da variável aleatória X, depende somente da distribuição de X. Portanto, em alguns casos podemos simplesmente nos referir a entropia de Shannon associada a uma distribuição de probabilidades, e denotar $H(p_i)$. Para que a entropia de Shannon faça sentido mesmo se $p_i = 0$ para algum i, definimos (veja

Exercício 2.1) $0 \log 0$ como sendo 0. Aqui log denota logaritmo na base 2, mais adequada para aplicações da entropia à teoria da informação, motivação original de Shannon.

Exercício 2.1 *Mostre que* $\lim_{x \to 0^+} x \ln x = 0$.

A grandeza $H(X)$ pode ser interpretada como uma medida da nossa incerteza sobre o valor da variável aleatória X, antes de observá-la, ou como a quantidade de informação sobre X que foi ganha depois de realizar a observação. Em particular, se X assume algum valor ω com probabilidade 1 então nenhuma informação é ganha após a observação de X (já que com probabilidade 1, X assume o valor ω). Neste caso, podemos verificar diretamente da definição de entropia que $H(X)$ é zero. Por outro lado, se X tem distribuição equiprovável, $p_i = 1/n$ para todo i, então $H(X)$ é exatamente $\log n$.

$H(X)$ é sempre não-negativa e limitada superiormente de maneira uniforme. Os exercícios a seguir mostram este resultado.

Exercício 2.2 *i) Mostre que podemos eliminar os valores de p identicamente nulos da soma (2.1) sem que se altere o valor de $H(X)$, isto é, $H(X) = -\sum_{\{i\,:\,p_i>0\}} p_i \log p_i$.*
ii) Usando o item i) e definindo $M(X) \equiv \max_{\{i\,:\,p_i>0\}} |\log p_i|$, mostre que

$$0 \leq H(X) \leq M(X).$$

Exemplo 2.1 (entropia binária) *Seja X uma variável de Bernoulli com parâmetro p. Então*

$$H(X) = -p \log p - (1-p) \log(1-p). \tag{2.2}$$

Exercício 2.3 *Mostre que a entropia binária, quando vista como função de p, tem um máximo em $p = 1/2$ e que seu valor máximo é 1. Conclua então que $H(X) \leq 1$ para qualquer v.a. X com distribuição binária.*

No exercício 2.3 fica clara a comodidade do logaritmo ser tomado na base 2. Com a escolha de outra base teríamos $\log 2$ como valor máximo.

Observe que a cota superior dada pelo Exercício 2.3 é uniforme em X, ao contrário da cota $M(X)$ do Exercício 2.2, que depende da distribuição de probabilidade de X. O Exercício 2.3 pode ser generalizado, como mostramos a seguir.

Exercício 2.4 (entropia maximal) *Use multiplicadores de Lagrange para concluir que o valor máximo de $H(X)$, como função das n variáveis $p_1, p_2, \cdots, p_n$, e restrita à condição $\sum_i p_i = 1$, $p_i \geq 0$, é igual a $\log n$. Mostre, também, que esse valor ocorre exatamente na distribuição equiprovável.*

Dos Exercícios 2.2 e 2.4 podemos concluir o seguinte teorema

Teorema 2.1 *Seja $X : \Omega \to \Lambda$ uma variável aleatória assumindo os valores discretos $\lambda_1, \lambda_2, \cdots, \lambda_n$ com probabilidades $p_1, p_2, \cdots, p_n$, respectivamente. Então*

$$0 \leq H(X) \leq \log n, \tag{2.3}$$

e o valor máximo é atingido quando $p_i = 1/n$ para todo i.

Novamente pode ser apreciada a conveniência da base 2: o menor inteiro igual ou superior a $\log n$ corresponde ao número de *bits* necessários para escrever o número n (ou seja, escrever o número n em base 2).

2.2 Concavidade de $H(X)$

Definição 2.2 *Um subconjunto C de um espaço vetorial V é* convexo *se, para todos u e $v \in C$, e todo $\alpha \in [0, 1]$, a* combinação convexa $\alpha u + (1 - \alpha)v \in C$.

Exercício 2.5 *Use indução para mostrar que, se C é um conjunto convexo e $\alpha_1, \cdots, \alpha_n \in [0, 1]$ tais que $\sum_{i=1}^n \alpha_i = 1$, então $\sum_{i=1}^n \alpha_i u_i \in C$ sempre que $u_i \in C$ para todo i.*

Definição 2.3 *Um elemento $u \in C$ é um* ponto extremal de C *se u não puder ser escrito como a combinação convexa de dois outros elementos de C.*

Exercício 2.6 *Mostre que $\{(p_1, p_2, \cdots, p_n) \in \mathbb{R}^n : p_i \geq 0 \text{ e } \sum_i p_i = 1\}$ é um conjunto convexo e determine os seus pontos extremais. Para $n = 2$ e $n = 3$, faça um desenho do conjunto e verifique que o ponto de máximo de $H(X)$ é o baricentro da figura obtida.*

Definição 2.4 *Uma função $f : C \to \mathbb{R}$ é* convexa *se o seu domínio C é convexo e se $f(\alpha u+(1-\alpha)v) \leq \alpha f(u)+(1-\alpha)f(v)$ para quaisquer $u, v \in C$ e para qualquer $\alpha \in [0,1]$.*

Definição 2.5 *Uma função $f : C \to \mathbb{R}$ é* estritamente convexa *se ela for convexa e se a igualdade $f(\alpha u + (1-\alpha)v) = \alpha f(u) + (1-\alpha)f(v)$ ocorrer somente quando $u = v$.*

Exercício 2.7 (Desigualdade de Jensen) *Seja $f : C \to \mathbb{R}$ uma função convexa. Então para todos $\alpha_1, \cdots, \alpha_n \in [0,1]$ tais que $\sum_i \alpha_i = 1$ e para todos $u_1, \cdots, u_n \in C$, $f(\sum_i \alpha_i u_i) \leq \sum_i \alpha_i f(u_i)$. Mostre ainda que, se f for estritamente convexa então a igualdade vale se, e somente se, $u_1 = u_2 = \cdots = u_n$.*

Exercício 2.8 *Definindo $\alpha_i = 1/n$ para todo i, definindo $\psi(p) = \ln p$ e observando que $[\sum_i \psi(p_i)]/n = \sum_i \alpha_i \psi(p_i)$, use a desigualdade de Jensen para obter uma outra prova do Teorema 2.1.*

Definição 2.6 *Dizemos que f é* côncava *se $-f$ é convexa.*

Exercício 2.9 *Mostre que a função $\varphi(t) : [0,\infty) \to \mathbb{R}^+$, definida como sendo zero na origem e sendo $-t \ln t$ nos outros pontos, é uma função côncava de t.*

Use os Exercícios 2.6 e 2.9 para mostrar que

Teorema 2.2 (Concavidade da Entropia) *$H(X)$, como função das variáveis $p_1, \cdots, p_n$, é uma função côncava no domínio $\{(p_1, \cdots, p_n) \in \mathbb{R}^n : p_i \geq 0$ e $\sum_i p_i = 1\}$.*

2.3 Entropia Relativa

Sejam X e Y duas variáveis aleatórias com distribuições $(p_1, p_2, \cdots, p_n)$ e $(q_1, q_2, \cdots, q_n)$, respectivamente.

Definição 2.7 *Definimos a* entropia relativa *de X com relação a Y como*

$$H(X\|Y) \equiv \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right). \tag{2.4}$$

Exercício 2.10 *Mostre que* $H(X\|Y) = -H(X) - \sum_i p_i \ln q_i$.

Exercício 2.11 *Se* $p_1 = 1$ *e* $p_2 = 0$, $q_1 = q > 0$ *e* $q_2 = 1-q$, *calcule* $H(X\|Y)$ *e* $H(Y\|X)$ *e verifique que* $H(X\|Y) \neq H(Y\|X)$.

A entropia relativa nos permitirá quantificar, num certo sentido, o quão distintas estão as duas distribuições de probabilidade $(p_i)_{i=1}^n$ e $(q_i)_{i=1}^n$. Contudo, a entropia relativa não poderia nos fornecer uma noção de distância pois, como o Exercício 2.11 nos mostra, ela não é simétrica. A entropia relativa, vista como uma medida da proximidade entre duas distribuições, será um conceito um pouco mais fraco do que o de distância. Contudo, a entropia relativa permeará resultados importantes no decorrer do texto.

Exercício 2.12 *Mostre que* $\ln x \leq x - 1$ *para todo* $x > 0$ *e que a igualdade vale se, e somente se,* $x = 1$.

Teorema 2.3 (Positividade da entropia relativa) *Sejam* X *e* Y *duas variáveis aleatórias com distribuições* $(p_i)_{i=1}^n$ *e* $(q_i)_{i=1}^n$, *respectivamente. Então* $H(X\|Y) \geq 0$ *e a igualdade ocorre se, e somente se,* $(p_i)_{i=1}^n = (q_i)_{i=1}^n$.

Prova: Da definição de entropia relativa e do o Exercício 2.12, nós obtemos

$$\begin{aligned} H(X\|Y) &= \textstyle\sum_i p_i \ln\left(\frac{p_i}{q_i}\right) \\ &= -\textstyle\sum_i p_i \ln\left(\frac{q_i}{p_i}\right) \\ &\geq \textstyle\sum_i p_i \left(1 - \frac{q_i}{p_i}\right) \quad = 0. \end{aligned}$$

É claro que se $p_i = q_i$ para todo i então $H(X\|Y) = 0$. Por outro lado, se $H(X\|Y) = 0$ então segue da seqüência de desigualdades acima que

$$0 = -\sum_i p_i \ln\left(\frac{q_i}{p_i}\right) \geq \sum_i p_i \left(1 - \frac{q_i}{p_i}\right) = 0.$$

Equivalemente

$$\sum_i p_i \left[\left(\frac{q_i}{p_i} - 1\right) - \ln\left(\frac{q_i}{p_i}\right)\right] = 0,$$

e usando o Exercício 2.12 nós concluímos que $p_i = q_i$ para todo i. □

Uma bela aplicação da entropia relativa é uma outra prova do Teorema 2.1, como o próximo exercício mostra.

Exercício 2.13 *Use a positividade da entropia relativa e o Exercício 2.10 com X tendo distribuição equiprovável para obter uma outra prova do Teorema 2.1.*

Usaremos freqüentemente esta técnica de encontrar expressões para quantidades entrópicas em termos da entropia relativa, tanto no contexto de informação clássica quanto quântica.

2.4 A Entropia Conjunta

Definição 2.8 (Distribuição Conjunta) *Sejam $X, Y : \Lambda \to \Omega$ variáveis aleatórias com distribuições $p_i = P(X = \omega_i)$ e $q_j = P(Y = \omega_j)$. A* distribuição conjunta *do vetor aleatório (X, Y) é caracterizada pelo vetor de probabilidades $p_{ij} = P(X = \omega_i, Y = \omega_j)$. Podemos verificar que $p_i = \sum_j p_{ij}$, $q_j = \sum_i p_{ij}$ e que quando X e Y são independentes $p_{ij} = p_i q_j$.*

Definição 2.9 (Entropia Conjunta) *Dado o vetor aleatório (X, Y), definimos sua* entropia conjunta *por*

$$H(X, Y) = -\sum_{ij} p_{ij} \log p_{ij}, \tag{2.5}$$

onde p_{ij} é a distribuição conjunta de (X, Y).

Exercício 2.14 *Mostre que, se X e Y são independentes, $H(X, Y) = H(X) + H(Y)$.*

O Exercício 2.14 diz que, se X e Y são independentes, uma variável não carrega informação sobre a outra, ou seja o total da informação no par é a soma das informações em cada uma. A seguir obteremos que em geral a entropia é subaditiva, ou seja, o total de informação no par é menor que a soma da informação em seus constituintes.

Teorema 2.4 (Subaditividade da Entropia de Shannon) *Sejam X e Y variáveis aleatórias com distribuições p_i e q_j, respectivamente. Então*

$$H(X, Y) \leq H(X) + H(Y),$$

com a igualdade sendo válida se, e somente se, X e Y são independentes.

Prova: Sejam $Z = (X, Y)$ e $W = (W_1, W_2)$ vetores aleatórios com distribuições conjuntas $(p_{ij})_{i,j}$ e $(p_i q_j)_{i,j}$, respectivamente (isto é, W é um vetor aleatório cuja distribuição conjunta seria a de Z caso X e Y fossem independentes). Segue então

$$\begin{aligned} H(Z\|W) &= \sum_{ij} p_{ij} \log \frac{p_{ij}}{p_i q_j} \\ &= \sum_{ij} p_{ij} \log p_{ij} - \sum_{ij} p_{ij} \log p_i - \sum_{ij} p_{ij} \log q_j \\ &= \sum_{ij} p_{ij} \log p_{ij} - \sum_{i} p_i \log p_i - \sum_{j} q_j \log q_j \\ &= -H(Z) + H(X) + H(Y). \end{aligned}$$

Pelo teorema 2.3 temos $H(Z\|W) \geq 0$ valendo a igualdade se, e só se, Z e W têm a mesma distribuição, ou seja, $p_{ij} = p_i q_j, \forall i, j$, logo X e Y são independentes. □

2.5 Entropia Condicional e Informação Mútua

A subaditividade sugere que pode haver *correlação* entre as variáveis, que ao aprender sobre uma podemos também ganhar informação sobre a outra. Uma pergunta natural então é saber quanta informação sobre X *não* está disponível em Y.

Definição 2.10 *A entropia condicional de X dado o conhecimento de Y é definida por*

$$H(X|Y) = H(X, Y) - H(Y). \tag{2.6}$$

Podemos interpretar a entropia condicional como a medida da incerteza que temos do valor de X dado que conhecemos o valor de Y. Pelo exercício 2.14 vemos que, se X e Y são independentes, $H(X|Y) = H(X)$, ou seja, o conhecimento de Y não traz informação sobre X.

Uma outra quantidade que podemos definir é a informação mútua contida no par X e Y, com objetivo de medirmos quanta informação elas têm em comum. Suponhamos que adicionemos a informação contida em X, $H(X)$, à informação contida em Y. A informação comum existente em X e Y será contada duas vezes nesta soma, enquanto a informação que não é comum será contada apenas uma vez. Se subtrairmos a informação conjunta de (X, Y), $H(X, Y)$, da soma feita anteriormente obteremos a informação comum ou a informação mútua de X e Y.

Definição 2.11 *A informação mútua contida em X e Y é definida por*

$$H(X:Y) \equiv H(X) + H(Y) - H(X,Y) \tag{2.7}$$

Note que segue diretamente das definições dadas acima que $H(X:Y) = H(X) - H(X|Y)$. E então temos uma importante igualdade relacionando a entropia condicional a informação mútua. É importante ressaltar que a informação mútua é simétrica, diferentemente da entropia condicional.

Exercício 2.15 *Seja (X,Y) um vetor aleatório tal que $X = f(Y)$.*
i) Mostre que $H(X|Y) = 0$;
ii) Mostre que $H(X:Y) = H(X)$;
iii) Enuncie e interprete resultados análogos a i) e ii) para o caso $Y = g(X)$.

2.6 Compactação de Dados e o Teorema de Shannon

Um problema importante, do ponto de vista prático, é o problema da compactação de dados. Originalmente este era um problema para as empresas de telefonia, preocupadas em transmitir voz e dados ocupando o mínimo possível o seu *canal* de transmissão. Hoje o problema faz parte do dia a dia de várias pessoas: ao usar um compactador para diminuir o tamanho de arquivos no computador, ou o sistema T9 de digitação de mensagens em telefones celulares, estamos utilizando a compactação da qual trata o teorema de Shannon.

A pergunta natural é: se temos determinada informação a ser armazenada ou transmitida, qual a melhor maneira de fazê-lo? De outra forma, pensemos no exemplo de uma fotografia: podemos dividir o tamanho da imagem em pedaços tão pequenos que nossos olhos não percebam que formam um reticulado, e assim armazenar a cor que deve ser utilizada em cada pedaço. Tais pedaços são os chamados *pixels* e esta é essencialmente a estratégia usada para gravar arquivos em formato bmp. Mas isso costuma ser muito custoso. Pense que sua foto tem um céu azul, areia branca e um coqueiro. Serão muitos pontos em azul, muitos pontos em branco, vários em marrom e outros tantos verdes. Não haveria uma maneira mais eficiente de guardar nossa imagem? Há, e há até mais de uma maneira, mas será que existe uma *melhor* maneira? É isso que o teorema de Shannon vai nos mostrar: não a melhor maneira, mas quantos bits precisarão ser registrados para gravar a imagem da maneira mais econômica possível.

Definição 2.12 *Considere uma seqüência de variáveis aleatórias independentes e identicamente distribuídas $X_1, \ldots, X_n$. Dado um $\varepsilon > 0$ dizemos que uma seqüência de valores $x_1, \ldots, x_n$ é ε-típica se*

$$2^{-n(H(X)+\varepsilon)} \leq P\left(X_1 = x_1, \ldots, X_n = x_n\right) \leq 2^{-n(H(X)-\varepsilon)}.$$

Será conveniente definirmos

$$T\left(n, \varepsilon\right) = \left\{(x_1, ..., x_n) \in \Omega^n : 2^{-n(H(X)+\varepsilon)} \leq P\left(X_1 = x_1, \ldots, X_n = x_n\right) \leq 2^{-n(H(X)-\varepsilon)}\right\},$$

ou seja o conjunto de todas as seqüências ε-típicas de comprimento n. Daqui até o fim do capítulo, todas as seqüências de variáveis aleatórias serão iid.

Exercício 2.16 *Para todo $n \in \mathbb{N}$ e $\varepsilon > 0$, mostre que:*
i) Se $P\left(X = \omega_i\right) = 1$ para algum i, então $\#T\left(n, \varepsilon\right) = 1$;
ii) Se a distribuição for equiprovável, então $\#T\left(n, \varepsilon\right) = \#\Omega^n$.
Descreva os resultados acima em palavras.

Teorema 2.5 (Seqüências Típicas)

i) Dado $\varepsilon > 0$, temos

$$\lim_{n\to\infty} P\left(T\left(n, \varepsilon\right)\right) = 1.$$

Em palavras, a probabilidade de uma seqüência de comprimento n ser ε-típica converge para 1 quando seu comprimento n vai para infinito.

ii) Dados $\varepsilon > 0$ e $\delta > 0$, existe $n_o \in \mathbb{N}$ tal que

$$\left(1 - \delta\right) 2^{n(H(X)-\varepsilon)} \leq \#T(n, \varepsilon) \leq 2^{n(H(X)+\varepsilon)}, \quad \forall n > n_o.$$

iii) Sejam $0 < R < H\left(X\right)$ número real e $S\left(n\right)$ um conjunto de seqüências de comprimento n tal que $\#S\left(n\right) \leq 2^{nR}$. Então, dado $\delta > 0$, existe $n_o \in \mathbb{N}$ tal que

$$\sum_{(x_1,\ldots,x_n)\in S(n)} P\left(X_i = x_i, i = 1, \ldots, n\right) < \delta, \quad \forall n > n_o.$$

Prova:

i) Dada a seqüência de variáveis aleatórias (X_i), defina a seqüência (Y_i) como $Y_i(\omega) = -\log p_j$ se $X_i(\omega) = x_j$. Neste caso, como a seqüência (X_i) é iid, a seqüência (Y_i) também o é. Logo, pela Lei dos Grandes Números (teorema 1.5) temos que $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^n Y_i}{n} - EY\right| \leq \varepsilon\right) = 1.$$

Observando que $EY = H(X)$ temos a seguinte igualdade entre os conjuntos

$$\begin{aligned}\left\{\left|\tfrac{\sum_{i=1}^n Y_i}{n} - EY\right| \leq \varepsilon\right\} &= \left\{n(H(X)-\varepsilon) \leq \textstyle\sum_{i=1}^n Y_i \leq n(H(X)+\varepsilon)\right\} \\ &= \left\{n(H(X)-\varepsilon) \leq -\log(p_1\cdots p_n) \leq n(H(X)+\varepsilon)\right\} \\ &= T(n,\varepsilon).\end{aligned}$$

ii) Como

$$1 \geq P(T(n,\varepsilon)) = \sum_{(x_1,\ldots,x_n)\in T(n,\varepsilon)} P(X_i = x_i) \geq (\#T(n,\varepsilon))\, 2^{-n(H(X)+\varepsilon)},$$

temos $\#T(n,\varepsilon) \leq 2^{n(H(X)+\varepsilon)}$, $\forall n \in \mathbb{N}$. Pelo item i), como $\lim_{n\to\infty} P(T(n,\varepsilon)) = 1$, dado $\delta > 0$, $\exists n_o \in \mathbb{N}$ tal que $P(T(n,\varepsilon)) > 1-\delta$, $\forall n > n_o$. Então

$$1 - \delta \leq P(T(n,\varepsilon)) \leq (\#T(n,\varepsilon))\, 2^{-n(H(X)-\varepsilon)},$$

logo $\#T(n,\varepsilon) \geq (1-\delta)\, 2^{n(H(X)-\varepsilon)}$.

iii) Dado $R < H(X)$, tome $\varepsilon > 0$ tal que $H(X) - R - \varepsilon > 0$. Denotando $S = S(n)$ e $T = T(n,\varepsilon)$, podemos escrever

$$\begin{aligned}&\sum_{(x_1,\ldots,x_n)\in S} P(X_i = x_i, i = 1,\ldots,n) = \\ &= \sum_{(x_1,\ldots,x_n)\in S\cap T} P(X_i = x_i, i = 1,\ldots,n) + \sum_{(x_1,\ldots,x_n)\in S\cap T^c} P(X_i = x_i, i = 1,\ldots,n).\end{aligned}$$

Pelo item i), dado $\varepsilon > 0$ existe $n_1 \in \mathbb{N}$ tal que

$$P(T(n,\varepsilon)) > 1 - \frac{\delta}{2}, \quad \forall n > n_1,$$

logo

$$\sum_{(x_1,\ldots,x_n)\in S\cap T^c} P(X_i = x_i, i = 1,\ldots,n) < \frac{\delta}{2}.$$

Por outro lado, pela própria definição de seqüência típica,

$$\sum_{(x_1,\ldots,x_n)\in S\cap T} P\left(X_i = x_i, i = 1,\ldots,n\right) \leq \left(\#S\left(n\right)\right) 2^{-n(H(X)-\varepsilon)} \leq 2^{-n(H(X)-R-\varepsilon)},$$

como $H\left(X\right)-R-\varepsilon > 0$, existe $n_2 \in \mathbb{N}$ tal que $2^{-n(H(X)-R-\varepsilon)} < \frac{\delta}{2}$, $\forall n > n_2$. Fazendo $n_o = \max\left\{n_1, n_2\right\}$ segue o resultado. □

Agora já temos todos os resultados que nos permitirão demonstrar o Teorema de Shannon. Falta apenas entender o que chamamos um sistema de Compressão-Descompressão, como usado por um computador para "zipar" um arquivo. Vamos manter o exemplo de um arquivo fotográfico em mente, embora o Teorema originalmente estivesse preocupado com a transmissão de dados. Nosso arquivo original contém a cor de cada pixel. Se usamos k bits para designar uma cor, e o reticulado possui m posições, devemos registrar $n = mk$ bits. Se considerarmos cada bit como uma variável aleatória iid[1], podemos pensar na entropia dessa distribuição. No nosso exemplo do coqueiro, há grande predomínio de poucas cores, o que diz que a entropia é pequena, comparada ao máximo valor que ela poderia assumir se tivéssemos muitas cores na foto. Um *sistema de compressão-descompressão com razão* R (com $0 < R < 1$) é um par de funções C e D, onde C leva n bits em Rn bits, enquanto D faz o caminho contrário. É claro que C e D não podem ser rigorosamente inversas, pois C não tem como ser injetora. Um sistema de compressão-descompressão será dito *confiável* se $P\left(D \circ Cx = x\right) > 1-\delta$, ou seja, se a possibilidade de falha puder ser controlada por um δ arbitrário. Em sua essência, o Teorema de Shannon diz que um sistema de compressão-descompressão precisa cuidar das seqüências típicas se quiser ser confiável.

Teorema 2.6 (Teorema de Shannon) *Suponha que* $\{X_i\}$ *seja uma seqüência iid de variáveis de Bernoulli com entropia* $H(X)$*. Suponha que* $R > H(X)$*. Então existe um esquema confiável de compressão com razão* R *para os dados desta fonte. Reciprocamente, se* $R < H(X)$ *qualquer esquema de compressão não é confiável.*

Prova: Suponha que $R > H(X)$. Escolha $\varepsilon > 0$ tal que $H(X) + \varepsilon < R$. Considere o conjunto $T(n, \varepsilon)$ das seqüências ε-típicas. Para qualquer $\delta > 0$ e suficientemente grande, existem no máximo $2^{n(H(X)+\varepsilon)} < 2^{nR}$ tais seqüências e a probabilidade da fonte produzir uma tal seqüência é no mínimo $1-\delta$. O método de compressão é simplesmente examinar agrupamentos de n dados de saída da fonte para ver se ele é ε-típico. Escolhida uma enumeração das típicas, associamos cada seqüência a seu valor; se a seqüência não for

[1] O que é uma simplificação, mas que pode ser melhor contornada.

típica, a associamos a um número específico entre 0 e nR que será para nós como um registro que indica falha. A descompressão destes tipos de seqüência será simplesmente uma outra seqüência gerada pela fonte aleatoriamente. E isto nos fornece um esquema eficiente de compressão de dados (pois quase toda seqüência é típica).
Supondo que $R < H(X)$. A combinação das operações compressão-descompressão tem no máximo 2^{nR} possíveis saídas, logo no maximo 2^{nR} das seqüências fornecidas pela fonte podem sofrer a operação citada acima sem a ocorrência de erro. Pelo teorema das seqüências típicas, para n suficientemente grande a probabilidade de uma seqüência fornecida pela fonte estar num conjunto de 2^{nR} seqüências tende a zero, se $R < H(X)$. Assim qualquer esquema de compressão não é confiável. □

Capítulo 3

Entropia e Teoria Quântica da Informação

Este capítulo pode ser visto como a versão quântica do capítulo anterior. Para isso é necessário introduzir os elementos centrais da teoria quântica, o que é feito em duas partes: sistemas quânticos simples na 3.1, compostos na 3.4. Novamente o resultado central é um teorema de compactação (o Teorema de Schumacher), mas assim como no capítulo anterior, várias relações de entropias são apresentadas, enfatizando especialmente o que reproduz e o que não reproduz a teoria clássica. A referência básica para aprofundamento é [5].

3.1 Noções rápidas de Mecânica Quântica

Não seria razoável, no espírito deste minicurso, nem assumir conhecimento prévio de Mecânica Quântica, nem querer abordá-la em detalhes. Vamos apresentar aqui apenas seus ingredientes necessários para a definição da *entropia de von Neumann*, a generalização natural do conceito de entropia para este cenário.

O passo mais importante que daremos é passar de um vetor de probabilidade para um *operador densidade*. Precisamos, porém, lembrar a definição de operadores positivos:

Definição 3.1 (Operador Positivo) *Seja V um espaço vetorial complexo, dotado de*

um produto escalar hermitiano. Um operador $O : V \to V$ é dito positivo *(semi-definido) se for auto-adjunto e, para todo $v \in V$, obedecer $(Ov, v) \geq 0$.*

Exercício 3.1 *Mostre que todos os autovalores de um operador positivo (semi-definido) são não-negativos.*
Dica: Todo operador auto-adjunto é diagonalizável.

Exercício 3.2 (Operadores Positivos Definidos) *O "semi-definido" refere-se à permissão que $(Ov, v) = 0$ sem qualquer restrição sobre v. Escreva uma definição para operador* positivo definido. *Depois adapte adequadamente o exercício 3.1 e, se possível, relacione com as noções equivalentes para formas quadráticas.*

Exercício 3.3 (Função de Operadores) *Se f é uma função analítica em $\mathbb{C}$, com série de potências $\sum_j a_j z^j$, podemos calcular formalmente esta função em um operador A usando $\sum_j a_j A^j$. Mostre que, se D for diagonalizável com decomposição espectral $D = \sum_k d_k P_k$, então*

$$f(D) = \sum_k f(d_k) P_k.$$

Em mecânica quântica, o espaço vetorial sobre o qual trabalharemos[1] chama-se *espaço de estados*. Um caso de particular interesse é o caso de dimensão 2, que será o análogo quântico da variável de Bernoulli. Em teoria de informação quântica, este importante exemplo é chamado um *qubit*.

A generalização adequada do conceito de probabilidade é dada então pelo seguinte objeto:

Definição 3.2 (Operador Densidade) *Um operador ρ é dito um* operador densidade *se for positivo (semi-definido) e obedecer* $\mathrm{Tr}\rho = 1$. *Um* estado *é caracterizado por seu operador densidade.*

Exercício 3.4 (Autovalores do Operador Densidade) *O que podemos falar sobre os autovalores de um operador densidade?*

[1] Da mesma forma que nos restringimos a espaços amostrais de cardinalidade finita, neste texto só abordaremos espaços de estados com dimensão finita.

Exercício 3.5 (Combinações Convexas) *Mostre que combinações covexas de operadores densidade são também operadores densidade. O que se conclui sobre o conjunto dos operadores densidade?*

Exercício 3.6 (Estados Puros) *Mostre que os pontos extremais do conjunto de operadores densidade correspondem a projetores ortogonais sobre subespaços unidimensionais de V. Tais pontos extremais são denominados* estados puros.

Em probabilidades, a distribuição de uma variável aleatória se refere às probabilidades de obter cada resultado possível em sorteios independentes dessa variável. É importante distinguir a noção de sorteio independente da realização de duas observações subseqüentes do mesmo sistema. Para descrever o resultado de sorteios da Mega-Sena, por exemplo, devemos acreditar que cada número é sorteado com a mesma probabilidade e independentemente, o que leva a todas as combinações de seis dezenas serem equiprováveis, excluídas as repetições proibidas pela regra do jogo. Assim, a maneira de descrevermos o resultado de um concurso ainda não realizado é através da distribuição equiprovável no conjunto de todas as combinações permitidas. O problema muda de figura quando tratamos de um sorteio já realizado. Se alguém perguntar "qual o resultado do concurso 600 da Mega-Sena?", e não tivermos informação alguma a esse respeito, ainda temos a distribuição equiprovável associada a ele. Mas se alguém nos informar que o resultado é $\{16, 18, 31, 34, 39, 54\}$, passamos a descrever tal variável "aleatória" por uma distribuição concentrada neste resultado agora conhecido. Qualquer nova observação da variável "concurso 600 da Mega-Sena" dará o mesmo resultado.

Em mecância quântica estas duas situações também aparecem. Em termos físicos, é comum dizer que um operador densidade caracteriza (o resultado de) um *processo de preparação*. Sistemas igualmente preparados são caracterizados pelo mesmo operador densidade, que permite associar a cada *processo de medição* a distribuição de uma variável aleatória. A definição a seguir nos ensina a fazer essa associação.

Definição 3.3 (Processo de Medição) *Sejam V o espaço de estados e ρ o operador densidade. Um* processo de medição *é dado por um conjunto de* operadores de medição $\{M_i\}_{i=1}^n$ *sobre V, com a propriedade*

$$\sum_{i=1}^{n} M_i^\dagger M_i = I, \tag{3.1}$$

onde I denota o operador identidade e $A^\dagger$ o adjunto de A. Os diferentes resultados do

processo serão indexados por i, e cada possível resultado tem probabilidade

$$p_i = \mathrm{Tr}\left(M_i \rho M_i^\dagger\right). \tag{3.2}$$

Exercício 3.7 *Mostre que a eq. (3.2) define uma distribuição de probabilidade.*

Exemplo 3.1 (Observáveis e medições projetivas) *Seja A um operador auto-adjunto sobre V. Pelo teorema da decomposição espectral,*

$$A = \sum_{i=1}^{n} a_i P_i,$$

com a_i reais e distintos, P_i projetores ortogonais com $P_i P_j = \delta_{ij} P_j$. O conjunto $\{P_i\}$ determina um processo de medição, com os resultados dados por $\{a_i\}$. O operador A é dito um observável. *O processo de medição aqui descrito é chamado* medição projetiva.

Da mesma forma que ao "concurso 600 da Mega-Sena" já sorteado, e cujo resultado já conhecemos, associamos uma variável aleatória com distribuição diferente daquela que o caracterizava antes do sorteio, em mecânica quântica associaremos um novo operador densidade a um sistema após sua medição.

Definição 3.4 *Sejam ρ operador densidade e $\{M_i\}_{i=1}^n$ operadores de medição. Se o resultado da medição for dado pelo índice i, o estado do sistema após a medição será dado por*

$$\rho_{após,i} = \frac{M_i \rho M_i^\dagger}{\mathrm{Tr}\left(M_i \rho M_i^\dagger\right)}. \tag{3.3}$$

Exercício 3.8 *Mostre que $\rho_{após,i}$ é um operador densidade. O que acontece com a definição 3.4 quando o denominador se anula?*

Exercício 3.9 *Que operador densidade devemos associar a um sistema sobre o qual foi realizado um processo de medição, mas cujo resultado não conhecemos? Como isso se compara com o caso de um sorteio da Mega-Sena já realizado, mas cujo resultado desconhecemos?*

3.1.1 Notação de Dirac

Existe uma notação bastante comum nos textos de mecância quântica, chamada *notação de Dirac*. Vamos introduzi-la aqui por dois motivos: primeiro para fazer uso subseqüente dela, segundo para que ela não se torne um obstáculo para aqueles que desejem ler outros textos sobre mecânica quântica. Tal notação é válida para espaços vetoriais V com produto escalar. Parece um tanto artificial, mas mostra-se de muita valia quando se opera com ela. Tudo começa com

Definição 3.5 *Um vetor $v \in V$ será denotado $|v\rangle$.*

Como temos um produto escalar em mãos, a cada $|v\rangle \in V$ podemos associar um *funcional linear* dado por "fazer produto escalar com $|v\rangle$". Tais funcionais constituem um espaço vetorial, chamado *dual* de V, e denotado V^*. Em notação de Dirac, temos:

Definição 3.6 *Denotamos $\langle v|$ o seguinte funcional linear:*

$$\begin{aligned} \langle v| : \ V &\longrightarrow \mathbb{C} \\ w &\longmapsto (v, w)\,. \end{aligned}$$

Vale realçar que o produto escalar que está sendo considerado deve ser linear na segunda variável, e semi-linear na primeira, ao contrário da maioria dos textos de álgebra linear. A motivação da notação vem justamente do fato que o produto escalar entre dois vetores é agora denotado $\langle v \mid w\rangle$, e como em inglês teríamos o *braket* dos vetores v e w, em notação de Dirac dizemos que $\langle v|$ é um "*bra*" (ou ainda, o bra v) e que $|w\rangle$ é um "*ket*" (o ket w).

Exemplo 3.2 (Projetores) *Seja $|v\rangle$ um vetor normalizado. O operador dado em notação de Dirac por $|v\rangle\langle v|$ é o* projetor ortogonal sobre o subespaço unidimensional gerado por $|v\rangle$*, usualmente chamado* projetor sobre $|v\rangle$*.*

Exercício 3.10 (Decomposição ortonormal) *Mostre que, dado um operador auto-adjunto A, podemos escolher uma base $\{|i\rangle\}_{i=1}^{d}$, tal que $A = \sum_i a_i |i\rangle\langle i|$.*

Exemplo 3.3 (Qubit em notação de Dirac) *No importante caso de um qubit, é comum denotar por $\{|0\rangle, |1\rangle\}$ uma base convenientemente escolhida. Esta base é chamada* base computacional*, em referência aos valores computacionais de um registrador binário.*

3.2 Entropia de von Neumann

A entropia de Shannon mede a incerteza associada a uma distribuição clássica de probabilidade. O conceito análogo em mecânica quântica será a entropia de von Neumann:

Definição 3.7 *Seja ρ operador densidade. Definimos a* entropia de von Neumann do estado associado a ρ *pela fórmula*

$$S(\rho) \equiv -\mathrm{Tr}\left(\rho \log \rho\right). \tag{3.4}$$

Na definição acima, foi feito uso do exercício 3.3, log é logaritmo na base 2, e novamente usamos $0 \log 0 \equiv 0$ (ver exercício 2.1).

Exercício 3.11 (Estados maximamente misturados) *Calcule a entropia de von Neumann para o operador I/d, onde I é o operador identidade e d a dimensão do espaço de estados. A nomenclatura ficará clara após o Teorema 3.2.*

Exercício 3.12 (Comparação entre entropias clássica e quântica) *Sejam*

$$\rho = p|0\rangle\langle 0| + (1-p)|1\rangle\langle 1| \quad e \quad \sigma = p|0\rangle\langle 0| + (1-p)\frac{(|0\rangle + |1\rangle)(\langle 0| + \langle 1|)}{2}.$$

Calcule $S(\rho), S(\sigma)$ e compare estes valores com a entropia binária para uma variável de Bernoulli com parâmetro p.

Assim como foi feito com a entropia de Shannon, definiremos uma versão quântica da entropia relativa.

Definição 3.8 *Suponha que ρ e σ são operadores densidade. A* entropia relativa de ρ com relação a σ *é definida por*

$$S(\rho\|\sigma) \equiv \mathrm{Tr}(\rho \log \rho) - \mathrm{Tr}(\rho \log \sigma). \tag{3.5}$$

Esta definição se completa com a convenção que a entropia relativa será $+\infty$ se o núcleo de σ tem interseção não-trivial com o suporte de ρ (o espaço vetorial gerado pelos autovetores de ρ associados aos autovalores não-nulos). Da mesma forma que no caso clássico, a entropia relativa é um conceito auxiliar, importante por suas propriedades que ajudam a demonstrar resultados de significado mais direto. Especial destaque deve ser dado ao teorema 3.1:

Teorema 3.1 (Desigualdade de Klein) *A entropia relativa quântica é não negativa.*

$$S(\rho\|\sigma) \geq 0, \tag{3.6}$$

com a igualdade ocorrendo se, e somente se, $\rho = \sigma$.

Prova: Sejam $\rho = \sum_i p_i|i\rangle\langle i|$ e $\sigma = \sum_j q_j|j\rangle\langle j|$ decomposições ortonormais (ver exercício 3.10) para ρ e σ. Pela definição de entropia relativa temos

$$S(\rho\|\sigma) = \sum_i p_i \log p_i - \sum_i \langle i|\rho \log \sigma|i\rangle. \tag{3.7}$$

Como $|i\rangle$ é autovetor de ρ, temos $\langle i|\rho = p_i\langle i|$ e

$$\langle i|\log\sigma|i\rangle = \langle i|\left(\sum_j \log q_j|j\rangle\langle j|\right)|i\rangle = \sum_j \log q_j P_{ij}, \tag{3.8}$$

onde $P_{ij} \equiv \langle i|j\rangle\langle j|i\rangle \geq 0$. A equação (3.7) pode ser reescrita como

$$S(\rho\|\sigma) = \sum_i p_i \left(\log p_i - \sum_j P_{ij} \log q_j\right). \tag{3.9}$$

Note que P_{ij} satisfaz[2] $P_{ij} \geq 0$, $\sum_i P_{ij} = 1$ e $\sum_j P_{ij} = 1$. Usando o fato de que o logaritmo é uma função estritamente côncava, da Desigualdade de Jensen obtemos a seguinte desigualdade, $\sum_j P_{ij} \log q_j \leq \log r_i$, onde $r_i \equiv \sum_j P_{ij} q_j$. E obtemos a igualdade se, e somente se, existe um valor j tal que $P_{ij} = 1$. □

Exercício 3.13 *Interprete a condição de existir um valor* j *tal que* $P_{ij} = 1$ *para concluir que vale a igualdade em*

$$S(\rho\|\sigma) \geq \sum_i p_i \log \frac{p_i}{r_i} \tag{3.10}$$

se, e somente se, P_{ij} *é uma* matriz de permutação.

Vamos relacionar como um teorema algumas propriedades importantes da entropia de von Neumann

Teorema 3.2 (Propriedades básicas da entropia de von Neumann)

[2]Matrizes satisfazendo essas três condições são chamadas duplamente estocásticas.

1. *A entropia de von Neumann é não-negativa. A entropia é zero se, e somente se, o estado é puro.*

2. *Num espaço $d-$dimensional a entropia é no máximo* $\log d$. *Este valor é atingido se, e somente se, o sistema é um* estado de mistura máxima, I/d.

3. *Suponha que p_i são probabilidades e que os operadores densidade ρ_i têm suporte em subespaços ortogonais. Então*

$$S\left(\sum_i p_i \rho_i\right) = H(p_i) + \sum_i p_i S(\rho_i). \tag{3.11}$$

Prova:
(1) Segue diretamente da definição.
(2) Da desigualdade de Klein (teorema 3.1), temos $0 \leq S(\rho \| I/d) = -S(\rho) + \log d$.
(3) Sejam λ_i^j e $|e_i^j\rangle$ autovalores e correspondentes autovetores de ρ_i. Observe que $p_i\lambda_i^j$ e $|e_i^j\rangle$ são autovalores e autovetores de $\sum_i p_i\rho_i$ e assim

$$\begin{array}{rcl} S\left(\sum_i p_i\rho_i\right) & = & -\sum_{ij} p_i\lambda_i^j \log p_i\lambda_i^j \\ & = & -\sum_i p_i \log p_i - \sum_i p_i \sum_j \lambda_i^j \log \lambda_i^j \\ & = & H(p_i) + \sum_i p_i S(\rho_i). \end{array}$$

□

Como interpretações, (1) novamente traz a interpretação de um quantificador das surpresas que podemos ter ao fazer medições, ou ainda, de quanta desordem há no estado ρ; (2) justifica que I/d seja o análogo quântico da distribuição equiprovável de probabilidades; (3) mostra como as entropias de Shannon e von Neumann se complementam - para melhor apreciar a importância da hipótese de suportes ortogonais, recomendamos voltar ao exercício 3.12.

3.3 Entropia e Medições

Classicamente, quando é realizado um sorteio de uma variável aleatória podemos considerar duas situações: quando ainda não sabemos o resultado, continuamos a descrevê-la pela mesma distribuição de probabilidade de antes do sorteio; quando ganhamos informação sobre o resultado (que pode ser o próprio resultado, ou apenas uma informação parcial), refazemos nossa descrição desta nova variável. Como ganhamos informação nesse processo, a nova distribuição tem menos entropia que a anterior. Quanticamente, uma

medição da qual não sabemos o resultado já modifica o sistema, como vimos no exercício 3.9. Além disso, pode haver medições que diminuem a entropia mesmo sem que saibamos seu resultado (exercício 3.14)!

Começamos então pela situação de medições projetivas das quais não sabemos o resultado:

Teorema 3.3 (Medições projetivas aumentam entropia) *Suponha que $\{P_i\}$ determina uma medição projetiva e que ρ é um operador densidade. Então a entropia do estado $\rho' = \sum_i P_i \rho P_i$, que descreve o sistema após a medição, obedece*

$$S(\rho') \geq S(\rho), \tag{3.12}$$

valendo a igualdade se, e só se, $\rho = \rho'$.

Prova: Pela desigualdade de Klein temos

$$0 \leq S(\rho \| \rho') = -S(\rho) - \mathrm{Tr}(\rho \log \rho'). \tag{3.13}$$

Usando que $\sum_i P_i = I$, $P_i^2 = P_i$, e a propriedade cíclica do traço, obtemos

$$\begin{aligned} -\mathrm{Tr}(\rho \log \rho') &= -\mathrm{Tr}\left(\textstyle\sum_i P_i \rho \log \rho'\right) \\ &= -\mathrm{Tr}\left(\textstyle\sum_i P_i \rho \log \rho' P_i\right). \end{aligned}$$

Como $\rho' P_i = P_i \rho P_i = P_i \rho'$, P_i comuta com $\log \rho'$. Segue que

$$\begin{aligned} -\mathrm{Tr}(\rho \log \rho') &= -\mathrm{Tr}\left(\textstyle\sum_i P_i \rho P_i \log \rho'\right) \\ &= -\mathrm{Tr}(\rho' \log \rho') = S(\rho'). \end{aligned}$$

A eq. (3.13) conclui a prova. □

O exercício seguinte mostra que medições generalizadas, das quais não sabemos o resultado, podem diminuir a entropia.

Exercício 3.14 *Suponha que um qubit num estado ρ é medido usando os operadores de medição, $M_1 = |0\rangle\langle 0|$ e $M_2 = |0\rangle\langle 1|$. Se o resultado da medição não é conhecido, após esta medição temos o sistema no estado $M_1 \rho M_1^\dagger + M_2 \rho M_2^\dagger$. Mostre que este procedimento pode diminuir a entropia do qubit. Como você interpretaria este processo?*

É natural pensar que ao ganharmos informação através de uma medição, a entropia do estado após a medição seja menor que a anterior. Mas será que essa "intuição clássica" vale no mundo quântico?

Exercício 3.15 (Medições Projetivas diminuem a Entropia) *Mostre que se* $\{P_i\}$ *determina uma medição projetiva e* ρ *o estado do sistema, então* $S(\rho_{i,após}) \leq S(\rho)$ *para todo* i*.*

Exercício 3.16 *Considere um qubit descrito pelo estado*

$$\rho = \frac{1}{2}|0\rangle\langle 0| + \frac{1}{4}(|0\rangle + |1\rangle)(\langle 0| + \langle 1|),$$

sujeito a um processo de medição dado por $M_1 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$ *e* $M_2 = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$*. O que acontece com a entropia de* ρ *nesse processo, para cada possível resultado?*

3.4 Sistemas quânticos compostos

Nosso objetivo é chegar ao análogo quântico do teorema de Shannon. Para isso deveremos considerar várias cópias de um sistema quântico. Precisamos antes aprender a tratar com *sistemas quânticos compostos*, ou seja, aqueles formados por várias partes.

O primeiro exemplo de sistema quântico composto são os *sistemas bipartites*. Seja A um sistema quântico descrito pelo espaço de estados V, e seja B outro sistema com espaço de estados W. Quando descrevemos conjuntamente os sistemas A e B (denotado AB), é claro que se $|v\rangle \in V$ e $|w\rangle \in W$ descrevem cada parte, o par $(|v\rangle, |w\rangle)$ descreverá o estado do sistema composto. Mas a linearidade da mecânica quântica exige que associemos um espaço de estados a cada sistema. Assim, devemos considerar o espaço vetorial gerado pelos pares $(|v\rangle, |w\rangle)$. Matematicamente este espaço é o *produto tensorial* $V \otimes W$, para o qual damos uma definição construtiva:

Definição 3.9 (Produto tensorial de espaços) *Sejam* V *e* W *espaços vetoriais (de dimensão finita) e* $\{|v_i\rangle\}_{i=1}^m$ *e* $\{|w_j\rangle\}_{j=1}^n$ *respectivas bases ortonormais. Construímos o espaço* $V \otimes W$ *declarando* $|v_i, w_j\rangle = (|v_i\rangle, |w_j\rangle)$ *uma base ortonormal desse espaço.*

Exercício 3.17 *Qual a dimensão de* $V \otimes W$*?*

Definição 3.10 (Produto tensorial de operadores) *Sejam* $A : V \to V$ *e* $B : W \to W$*. O produto tensorial de* A *e* B*,* $A \otimes B : V \otimes W \to V \otimes W$*, é definido na base produto por* $A \otimes B\,|v_i, w_j\rangle = A\,|v_i\rangle \otimes B\,|v_j\rangle$ *e estendido por linearidade.*

Como o estado de um sistema é caracterizado por um operador densidade ρ, que pela definição 3.3 prevê (probabilisticamente) os resultados de qualquer medição sobre o sistema, uma questão torna-se natural: como podemos descrever o estado de uma parte do sistema, a partir da descrição do sistema global? Descrever o sistema global é ter ρ^{AB}, um operador densidade sobre o espaço $V \otimes W$. Descrever uma parte (a parte A, por exemplo) significa obter um operador densidade ρ^A (sobre V) capaz de prever de maneira semelhante a ρ^{AB} os resultados de qualquer *medição local* realizada nesta parte. Comecemos então caracterizando um processo de medição local:

Definição 3.11 *Seja $\{M_i\}$ um processo de medição sobre o sistema A. Os operadores $\{M_i \otimes I_B\}$ constituem um* processo de medição local *(na parte A) sobre o sistema AB.*

Para entender esta nomenclatura é importante citar que um exemplo importante de sistema bipartite é composto por dois laboratórios "distantes" (onde distante pode significar duas mesas em uma mesma sala ou realmente quilômetros de distância, como já há experimentos feitos). Assim, um processo de medição local age apenas "na sua parte" do sistema.

Vamos agora à receita para obter ρ^A a partir de ρ^{AB}:

Definição 3.12 (Traço parcial) *Seja T um operador sobre $V \otimes W$. Escrevendo-o com respeito a uma base produto, temos*

$$T = \sum_{i,i'} \sum_{j,j'} |v_i, w_j\rangle \, T_{ij,i'j'} \, \langle v_{i'}, w_{j'}| \, .$$

O traço parcial de T com respeito à segunda componente *é o operador*

$$T^A \equiv \mathrm{Tr}_B T = \sum_{i,i'} |v_i\rangle \sum_j T_{ij,i'j} \, \langle v_{i'}| \, .$$

De maneira análoga define-se o traço parcial com respeito à primeira componente.

Exercício 3.18 *Mostre que as medidas de probabilidade geradas por $\left(\{M_i \otimes I_B\}, \rho^{AB}\right)$ e por $\left(\{M_i\}, \rho^A\right)$ usando as definições 3.3 e 3.12 são idênticas.*

Exercício 3.19 *Se $T = A \otimes B$, calcule T^A e T^B. O que acontece quando T é um operador densidade?*

Exercício 3.20 *Considere agora o sistema de* dois qubits. *Denotando a base produto por* $\{|00\rangle, |01\rangle, |10\rangle, |11\rangle\}$, *obtenha* ρ^A *e* ρ^B *para os seguintes estados:*

1. $\rho = |\psi\rangle\langle\psi|$, *onde* $|\psi\rangle = \frac{1}{\sqrt{2}}(|10\rangle + |11\rangle)$;
2. $\rho = |\Phi\rangle\langle\Phi|$, *onde* $|\Phi\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$;
3. $\rho = I/4$.

As definições apresentadas até aqui se generalizam imediatamente para *sistemas multipartites*. Suas conseqüências não necessariamente. Vamos agora discutir um resultado interessante de sistemas bipartites (mais precisamente, de produto tensorial de dois espaços de dimensão finita), que não admite generalização para sistemas com mais partes. Trata-se da *decomposição de Schmidt*.

Teorema 3.4 (Decomposição de Schmidt) *Seja* $V = V^A \otimes V^B$ *espaço vetorial e* $|\Psi\rangle \in V$. *Então existem bases ortonormais* $\{|a_i\rangle\}$ *para* V^A *e* $\{|b_j\rangle\}$ *para* V_b *tais que*

$$|\Psi\rangle = \sum_k c_k |a_k\rangle \otimes |b_k\rangle, \tag{3.14}$$

com a soma se estendendo até a dimensão do menor subespaço.

Prova: A demonstração vai usar o exercício 3.22, proposto abaixo, e por simplicidade demonstraremos apenas o caso de dimensões iguais para as duas partes. Comecemos escrevendo $|\Psi\rangle = \sum_{ij} m_{ij} |i\rangle \otimes |j\rangle$ com respeito a bases arbitrárias (ortonormais) $\{|i\rangle\}$ de V^A e $\{|j\rangle\}$ de V^B. Os coeficientes m_{ij} podem ser vistos como elementos de uma matriz quadrada M, que pela Decomposição a valor singular (exercício 3.22) pode ser escrita como $M = UDV$, com U e V unitárias e D diagonal com elementos reais e não-negativos. Escrevendo isso para os elementos de matriz,

$$m_{ij} = \sum_k u_{ik} d_k v_{kj},$$

que permite escrever o estado como

$$|\Psi\rangle = \sum_{ijk} u_{ik} d_k v_{kj} |i\rangle \otimes |j\rangle = \sum_k d_k \left(\sum_i u_{ik} |i\rangle\right) \otimes \left(\sum_j v_{kj} |j\rangle\right),$$

que tem a forma da eq. (3.14) com $|a_k\rangle = \sum_i u_{ik} |i\rangle$ e $|b_j\rangle = \sum_j v_{kj} |j\rangle$, e a ortogonalidade segue da unitariedade de U e V. □

Exercício 3.21 (Decomposição Polar) *Seja A um operador em V. Então $A^\dagger A$ é um operador positivo, e podemos definir $J = \sqrt{A^\dagger A}$. Mostre que existe U unitária tal que $A = UJ$, e que U é única se, e só se, A é invertível. Esta é chamada (uma)* decomposição polar à esquerda *do operador A. Compare com a decomposição polar de números complexos para entender a nomenclatura. Além disso, mostre a existência de uma* decomposição polar à direita.
Sugestão: considere $K = \sqrt{AA^\dagger}$.

Exercício 3.22 (Decomposição a valor singular) *Seja M matriz quadrada. Mostre que existem U e V unitárias e D diagonal com elementos não-negativos tais que*

$$M = UDV.$$

Sugestão: considere uma decomposição polar e diagonalize o operador positivo.

Exercício 3.23 (Decomposição de Schmidt geral) *Escreva a demonstração do caso geral da Decomposição de Schmidt.*

Uma interessante conseqüência da Decomposição de Schmidt é a idéia de *purificação*. Se $\rho = \sum_i p_i |i\rangle \langle i|$ é um operador densidade de um estado misto sobre um espaço V, podemos considerar um outro espaço auxiliar V' e o estado puro $|\Psi\rangle = \sum_i \sqrt{p_i} |i\rangle \otimes |i'\rangle$ em $V \otimes V'$, tal que, quando o segundo subespaço é ignorado, reobtemos o estado de mistura de onde começamos. Esse processo é chamado uma *purificação* de ρ, e mostra-se bastante útil para várias definições e demonstrações.

3.5 Entropia em sistemas compostos

Um primeiro resultado, bastante importante, é conseqüência direta da Decomposição de Schmidt e do fato da entropia de von Neumann só depender dos autovalores de ρ:

Teorema 3.5 *Se AB é um sistema composto descrito por um estado puro $\rho^{AB} = |\Psi\rangle \langle\Psi|$, então $S\left(\rho^A\right) = S\left(\rho^B\right)$.*

Outra propriedade interessante segue do item (3) do teorema 3.2:

Teorema 3.6 (Entropia Conjunta) *Suponha que p_i são probabilidades, $|i\rangle$ são estados ortogonais para um sistema A, e ρ_i qualquer conjunto de operadores para um outro sistema B. Então*

$$S\left(\sum_i p_i |i\rangle\langle i| \otimes \rho_i\right) = H(p_i) + \sum_i p_i S(\rho_i). \tag{3.15}$$

Exercício 3.24 (Entropia de um produto tensorial) *Use o teorema da entropia conjunta, para mostrar que $S(\rho \otimes \sigma) = S(\rho) + S(\sigma)$. Prove também este resultado diretamente da definição de entropia de von Neumann.*

Em analogia com a entropia de Shannon é possível definir no contexto de entropia de von Neumann as entropias conjunta e condicional e também a informação mútua para sistemas quânticos compostos. A entropia conjunta (do estado ρ^{AB}) de um sistema AB será dada por $S(A,B) \equiv S\left(\rho^{AB}\right) = -\mathrm{Tr}(\rho^{AB}\log(\rho^{AB}))$. Daí definimos, respectivamente, a entropia condicional e a informação mútua por:

$$S(A|B) \equiv S(A,B) - S(B) = S\left(\rho^{AB}\right) - S\left(\rho^{B}\right), \tag{3.16}$$

$$S(A:B) \equiv S(A) + S(B) - S(A,B) = S\left(\rho^{A}\right) + S\left(\rho^{B}\right) - S\left(\rho^{AB}\right). \tag{3.17}$$

Exercício 3.25 *Calcule $S(A|B)$ e $S(A:B)$ para os estados do exercício 3.20. Compare com as propriedades de $H(A|B)$ e $H(A:B)$ estudadas no capítulo anterior. Tente interpretar seu resultado.*

Deve ser claro então que buscar cotas e relações envolvendo estas diversas entropias nos permite entendê-las melhor. Um bom exemplo está aqui:

Teorema 3.7 (Subaditividade) *Sejam A e B dois sistemas quânticos distintos descritos pelo estado conjunto ρ^{AB}. Então a entropia conjunta deste sistema satisfaz a desigualdade*

$$S(A,B) \leq S(A) + S(B), \tag{3.18}$$

com a igualdade valendo apenas se $\rho^{AB} = \rho^A \otimes \rho^B$.

Prova: Pela desigualdade de Klein temos, $S(\rho) \leq -\mathrm{Tr}(\rho \log \sigma)$. Definindo $\rho \equiv \rho^{AB}$ e $\sigma = \rho^A \otimes \rho^B$ temos as seguintes identidades

$$\begin{aligned} -\mathrm{Tr}(\rho \log \sigma) &= -\mathrm{Tr}(\rho^{AB}(\log \rho^A + \log \rho^B)) \\ &= -\mathrm{Tr}(\rho^A \log \rho^A) - \mathrm{Tr}(\rho^B \log \rho^B) \\ &= S(A) + S(B). \end{aligned}$$

Usando então a desigualdade de Klein temos $S(A,B) \leq S(A) + S(B)$. Da condição de igualdade de desigualdade de Klein, segue a última afirmação. □

Queremos agora mostrar que a entropia de von Neumann é uma função côncava no conjunto dos operadores densidade, ou seja, dada uma distribuição de probabilidades p_i e correspondentes operadores densidade ρ_i, a entropia satisfaz a desigualdade

$$S\left(\sum_i p_i \rho_i\right) \geq \sum_i p_i S(\rho_i). \tag{3.19}$$

Para isso, suponhamos que ρ_i são estados de um sistema A e consideremos um sistema auxiliar B cujo espaço de estados tenha uma base ortonormal $\{|i\rangle\}$. Consideremos o estado

$$\rho^{AB} \equiv \sum_i p_i \rho_i \otimes |i\rangle\langle i|$$

para o sistema composto. Obtemos as seguintes identidades

$$\begin{array}{lcl} S(A) & = & S\left(\sum_i p_i \rho_i\right), \\ S(B) & = & S\left(\sum_i p_i |i\rangle\langle i|\right) = H(p_i), \\ S(A,B) & = & H(p_i) + \sum_i p_i S(\rho_i). \end{array}$$

Finalmente, usando a subaditividade de $S(A,B)$, temos (3.19).

Exercício 3.26 *Prove que vale a igualdade em (3.19) se, e somente se, a combinação convexa for trivial.*

Exercício 3.27 *Mostre que existe um conjunto de matrizes unitárias $\{U_j\}$ e um vetor de probabilidade (p_j) tais que, para qualquer matriz A,*

$$\sum_i p_i U_i A U_i^\dagger = \mathrm{Tr}(A)\frac{I}{d},$$

onde d é a dimensão do espaço onde A esta definida. Use esta informação e a concavidade estrita da entropia para dar outra demonstração de que a entropia atinge seu máximo na matriz $\frac{I}{d}$.

Exercício 3.28 *Seja P um projetor e $Q = I - P$ o projetor complementar. Prove que existem operadores unitários U_1 e U_2 e um parâmetro $p \in [0,1]$ tais que para todo ρ, $P\rho P + Q\rho Q = pU_1\rho U_1^\dagger + (1-p)U_2\rho U_2^\dagger$. Use esta observação para dar uma prova alternativa do teorema 3.3.*

3.6 Entropia de uma Mistura de Estados Quânticos

Por vezes o estudo de sistemas compostos ajuda a entender os sistemas simples. Aqui faremos isso: para demonstrar uma propriedade de um sistema simples, vamos considerá-lo como parte de um sistema composto. O resultado que queremos generaliza o teorema 3.6. Uma maneira interessante de interpretar o estado aqui trabalhado é que tenhamos um conjunto de possíveis preparações de estado quântico, cada qual caracterizada por seu operador densidade ρ_i, e um vetor de probabilidade (p_i) que define a chance de cada estado ser preparado. O estado de tal sistema será caracterizado por

$$\rho = \sum_i p_i \rho_i. \tag{3.20}$$

Vamos relacionar a entropia de tal estado com a entropia de von Neumann de cada constituinte e a entropia de Shannon do vetor (p_i).

Teorema 3.8 *Nas condições acima,*

$$S(\rho) \leq \sum_i p_i S(\rho_i) + H(p_i),$$

com igualdade se, e só se, os estados ρ_i têm suportes em subespaços dois a dois ortogonais.

Prova: Vamos supor inicialmente que cada ρ_i é estado puro, isto é, $\rho_i = |\psi_i\rangle\langle\psi_i|$, e que são estados de um sistema A. Vamos introduzir um sistema auxiliar B com uma base ortonormal $\{|i\rangle\}$ correspondente ao índice i de p_i. Defina

$$|AB\rangle \equiv \sum_i \sqrt{p_i}|\psi_i\rangle \otimes |i\rangle.$$

Como $|AB\rangle$ é um estado puro temos

$$S\left(\rho^B\right) = S\left(\rho^A\right) = S\left(\sum_i p_i |\psi_i\rangle\langle\psi_i|\right) = S(\rho).$$

Suponha que realizamos uma medição projetiva no sistema B, com operadores de medição $\{|i\rangle\}$, mas sobre a qual não sabemos o resultado. Após esta medição o estado do sistema B será

$$\rho'^B = \sum_i p_i |i\rangle\langle i|.$$

Mas pelo teorema 3.3 a entropia não decresce se realizamos medições projetivas (das quais não sabemos o resultado), logo $S(\rho) \leq S(\rho') = H(p_i)$. Observando que neste caso $S(\rho_i) = 0$ conclui-se que, quando cada ρ_i é um estado puro,

$$S(\rho) \leq H(p_i) = H(p_i) + \sum_i p_i S(\rho_i).$$

Além do mais a igualdade vale se, e somente se, $\rho^B = \rho'^B$ que é equivalente a dizer que os estados $|\psi_i\rangle$ são ortogonais.
Para provarmos o teorema no caso de misturas, consideremos $\rho_i = \sum_j p_j^i |e_j^i\rangle\langle e_j^i|$ decomposições ortonormais de cada estado ρ_i. Então $\rho = \sum_{ij} p_i p_j^i |e_j^i\rangle\langle e_j^i|$. Aplicando o resultado obtido para estados puros e usando o fato de que $\sum_j p_j^i = 1$, para cada i, temos

$$\begin{array}{rcl} S(\rho) & \leq & -\sum_{ij} p_i p_j^i \log(p_i p_j^i) \\ & = & -\sum_i p_i \log(p_i) - \sum_i p_i \sum_j p_j^i \log(p_j^i) \\ & = & H(p_i) + \sum_i p_i S(\rho_i) \end{array}$$

que estabelece a desigualdade. A condição para que valha a igualdade é estabelecida da mesma maneira que foi feito para o caso de estados puros. □

Intuitivamente a cota superior apresentada para a entropia de ρ, nos diz que a incerteza sobre o estado $\sum_i p_i \rho_i$ nunca é maior que a nossa incerteza média sobre o estado ρ_i mais uma contribuição $H(p_i)$ que representa a contribuição máxima possível para a nossa incerteza a respeito do índice i.

3.7 Teorema de Schumacher

Queremos agora obter o análogo quântico do Teorema de Shannon. Lembremos que os ingredientes essenciais lá (na secção 2.6) foram variáveis aleatórias iid, a noção de seqüência típica e a demonstração que um processo de compactação será *confiável* se, e só se, der conta das seqüências típicas de um tamanho fixado.

Nosso primeiro passo aqui será traduzir a noção de n variáveis aleatórias iid para o contexto quântico. Por simplicidade, consideremos apenas *qubits*. O estado de n qubits é caracterizado por um operador densidade em $H = V^{\otimes n}$, uma notação conveniente para o produto tensorial de n cópias do espaço V. Por serem qubits, $\dim V = 2$ e $\dim H = 2^n$. Mas só teremos n qubits iid se o operador ρ sobre H for da forma $\rho_1^{\otimes n}$, onde ρ_1 denota o estado de um qubit, e $A^{\otimes n}$ denota o produto tensorial de n cópias do operador A.

Exercício 3.29 *Seja $\rho = \rho_1^{\otimes n}$ como acima. Faça o traço parcial em $n-1$ qubits para obter o estado do qubit restante. Use seu resultado para justificar a afirmativa que esta é a generalização natural de iid.*

Vamos introduzir agora as idéias da versão quântica das seqüências típicas. Suponha que o operador densidade ρ_1 tenha uma decomposição ortonormal

$$\rho_1 = \sum_x p(x)|x\rangle\langle x|, \tag{3.21}$$

onde $\{|x\rangle\}$ é um conjunto ortonormal e $(p(x))$ um vetor de probabilidades (no caso de qubits, $x = 0, 1$, mas a definição pode considerar dimensão arbitrária; é importante notar que os vetores $|x\rangle$ são determinados pela diagonalização de ρ_1). Podemos definir sequências ε-típicas $\{x_1, \ldots, x_n\}$ como sendo as seqüências para as quais se verifica a desigualdade

$$\left| \frac{1}{n} \log \left(\frac{1}{p(x_1) \ldots p(x_n)} \right) - S(\rho) \right| \leq \varepsilon,$$

onde a única diferença para o caso clássico é a entropia utilizada. Um *estado ε-típico* é um estado $|x_1\rangle|x_2\rangle \ldots |x_n\rangle$ para o qual a seqüência $\{x_1, \ldots, x_n\}$ é ε-típica. Definimos o *subespaço ε-típico* como sendo o subespaço gerado por todos os estados ε-típicos. Denotaremos o subespaço ε-típico por $T(n, \varepsilon)$ e o projetor neste subespaço por $P(n, \varepsilon)$. Note que

$$P(n, \varepsilon) = \sum_{\{x_i\}_{i=1}^n \ \varepsilon\text{-típica}} |x_1\rangle\langle x_1| \otimes |x_2\rangle\langle x_2| \otimes \ldots \otimes |x_n\rangle\langle x_n|.$$

Teorema 3.9 (Teorema do Subespaço Típico) *Seja ρ operador densidade sobre V, com $P(n, \varepsilon)$ e $T(n, \varepsilon)$ definidos acima.*

1. *Dado $\varepsilon > 0$, temos*
$$\lim_{n \to \infty} \mathrm{Tr}(P(n, \varepsilon)\rho^{\otimes n}) = 1.$$
Em palavras, o traço da restrição de n cópias de ρ ao subespaço ε-típico tende a 1 quando n vai para infinito.

2. *Dados $\varepsilon > 0$ e $\delta > 0$, existe $n_o \in \mathbb{N}$ tal que*
$$(1 - \delta)2^{n(S(\rho) - \varepsilon)} \leq \dim(T(n, \varepsilon)) \leq 2^{n(S(\rho) + \varepsilon)}, \quad \forall n > n_o.$$

3. *Sejam $0 < R < H(\rho)$ um número real e $\Pi(n)$ um projetor sobre um subespaço de $V^{\otimes n}$ com $\dim(\Pi(n)) \leq 2^{nR}$. Então, dado $\delta < 0$, existe $n_o \in \mathbb{N}$ tal que*
$$\mathrm{Tr}(\Pi(n)\rho^{\otimes n}) < \delta, \quad \forall n > n_o.$$

Prova: Para provar (1) basta observar que

$$\mathrm{Tr}(P(n,\varepsilon)\rho^{\otimes n}) = \sum_{\{x_i\}_{i=1}^n \ \varepsilon\text{-típica}} p(x_1)\dots p(x_n)$$

e usar o primeiro item do Teorema 2.5.
(2) Segue diretamente do segundo item do Teorema 2.5.
(3) Vamos fazer o traço separadamente no subespaço ε-típico e no seu complemento ortogonal,

$$\mathrm{Tr}(\Pi(n)\rho^{\otimes n}) = \mathrm{Tr}(\Pi(n)\rho^{\otimes n}P(n,\varepsilon)) + \mathrm{Tr}(\Pi(n)\rho^{\otimes n}(I - P(n,\varepsilon))),$$

e cotar cada termo. Para o primeiro termo observe que

$$\rho^{\otimes n}P(n,\varepsilon) = P(n,\varepsilon)\rho^{\otimes n}P(n,\varepsilon),$$

já que $P(n,\varepsilon)$ é um projetor que comuta com $\rho^{\otimes n}$. Mas

$$\mathrm{Tr}(\Pi(n)P(n,\varepsilon)\rho^{\otimes n}P(n,\varepsilon)) \leq 2^{nR}2^{-n(S(\rho)-\varepsilon)},$$

já que os autovalores de $P(n,\varepsilon)\rho^{\otimes n}P(n,\varepsilon)$ são limitados superiormente por $2^{-n(S(\rho)-\varepsilon)}$. Assim, existe n_1 tal que

$$\mathrm{Tr}\left(\Pi\left(n\right)\rho^{\otimes n}P\left(n,\varepsilon\right)\right) \leq \frac{\delta}{2}, \quad \forall n > n_1.$$

Para o segundo termo note que $I - \Pi\left(n\right)$ é operador positivo. Já que $\Pi(n)$ e $\rho^{\otimes n}(I - P(n,\varepsilon))$ são ambos operadores positivos segue que

$$0 \leq \mathrm{Tr}\left(\Pi(n)\rho^{\otimes n}(I - P(n,\varepsilon)\right) \leq \mathrm{Tr}(\rho^{\otimes n}(I - P(n,\varepsilon)),$$

e pelo item (1) exite n_2 tal que

$$\mathrm{Tr}\left(\Pi(n)\rho^{\otimes n}(I - P(n,\varepsilon))\right) \leq \frac{\delta}{2}, \quad \forall n > n_2.$$

Basta tomar $n_o > \max\{n_1, n_2\}$ para demonstrar o resultado. □

Dos ingredientes básicos, ainda precisamos definir um *processo de compressão-descompressão* e buscar a noção adequada de *confiável*. Um *processo de compressão* (de n cópias do estado, ou ainda, de um "bloco" de tamanho n) será uma aplicação $\mathcal{C}_n : V^{\otimes n} \to V_c^n$, onde V_c^n é o chamado *espaço de compressão*. Um *processo de descompressão*v (para n cópias) é uma aplicação $\mathcal{D}_n : V_c^n \to V^{\otimes n}$. Um *processo de compressão-descompressão* é uma seqüência de processos $\{\mathcal{C}_n, \mathcal{D}_n\}_{n=1}^{\infty}$. A *taxa de compressão* do processo é dado pelo

limite $\lim_{n\to\infty} \frac{\log \dim(V_c^n)}{n}$, com a mesma interpretação que no caso clássico: a razão entre a quantidade de qubits necessária para armazenar de maneira confiável a informação contida em n qubits caracterizados pelo estado ρ e esse número total de qubits, n. É importante enfatizar que em aplicações práticas simplesmente toma-se n "suficientemente grande", mas na sua descrição teórica aparecem as seqüências e os limites acima.

Por fim, confiabilidade. A idéia deve ser clara: queremos que $\mathcal{D}_n \circ \mathcal{C}_n$ seja "efetivamente" próxima da identidade. É claro que se $\dim(V_c) < 2^n$, tal composição não pode ser rigorosamente a identidade. Mas não precisamos disso. O que precisamos é ter

$$\mathcal{D}_n \circ \mathcal{C}_n \rho^{\otimes n} \approx \rho^{\otimes n}.$$

Mas ainda falta dizer o que significa "$\approx$" nessa expressão. Para isso, usaremos a noção de *fidelidade de um processo quântico*. Inspirado nos *processos de medição*, podemos definir:

Definição 3.13 (Processos Quânticos) *Um* processo quântico $\mathcal{E}$ *será definido por um conjunto de operadores* $\{E_i\}$ *tais que* $\sum_i \mathrm{Tr}\left(E_i^\dagger E_i\right) = 1$ *pela seguinte operação*

$$\rho \longmapsto \sum_i E_i \rho E_i. \tag{3.22}$$

Definição 3.14 (Fidelidade) *Se um processo quântico* $\mathcal{E}$ *é dado pelos operadores* $\{E_i\}$*, a* fidelidade de $\mathcal{E}$ quando aplicado ao estado ρ *é dada por*

$$F(\mathcal{E}, \rho) = \sum_i \left|\mathrm{Tr}\left(\rho E_i\right)\right|^2. \tag{3.23}$$

Agora sim temos os ingredientes para a noção de confiável: para $\delta > 0$, um par $(\mathcal{C}_n, \mathcal{D}_n)$ é dito δ-confiável se

$$F\left(\mathcal{D}_n \circ \mathcal{C}_n, \rho^{\otimes n}\right) > 1 - \delta.$$

Dizemos que um processo de compressão é confiável se, dado $\delta > 0$, existe n_o tal que para todo $n > n_o$ os pares $(\mathcal{C}_n, \mathcal{D}_n)$ são δ-confiáveis.

Teorema 3.10 (Teorema de Schumacher) *Seja* ρ *um operador densidade em* V*. Se* $R > S(\rho)$ *então existe um esquema de compressão confiável de razão* R*. Se* $R < S(\rho)$ *então qualquer sistema de compressão de razão* R *não é confiável.*

Prova: Suponha $R > S(\rho)$ e seja $\varepsilon > 0$ tal que $S(\rho)+\varepsilon \leq R$. Pelo teorema dos subespaços típicos, para qualquer $\delta > 0$ e para todo n suficientemente grande, $\mathrm{Tr}(\rho^{\otimes n} P(n,\varepsilon)) \geq 1-\frac{\delta}{2}$ e $\dim(T(n,\varepsilon)) \leq 2^{nR}$. Nessa demonstração vamos considerar o espaço de codificação V_c^n como um subespaço de $V^{\otimes n}$, apenas para simplificar a notação. Escolha para V_c^n qualquer subespaço de $V^{\otimes n}$ de dimensão 2^{nR} contendo $T(n,\varepsilon)$ e construa o processo de compressão da seguinte maneira: primeiro realiza-se uma medição no conjunto de n qubits descrita pelos projetores ortogonais $\{P(n,\varepsilon), I - P(n,\varepsilon)\}$ e que retorna os valores 0 ou 1 respectivamente. Se o resultado é 0, nada é feito e o estado é deixado no subespaço típico. Se o resultado for 1, trocamos o estado do sistema por algum estado padrão "$|0\rangle$" escolhido no subespaço típico. Vemos que tal procedimento pode ser assim descrito:

$$\mathcal{C}^n(\sigma) \equiv P(n,\varepsilon)\sigma P(n,\varepsilon) + \sum_i A_i \sigma A_i^\dagger,$$

onde $A_i = |0\rangle\langle i|$ e $\{|i\rangle\}_{i=1}^N$ é uma base ortonormal para o complemento ortogonal do subespaço típico (N sua dimensão). O processo de descompressão será dado pela inclusão $\mathcal{D}_n : V_c^n \to V^{\otimes n}$, definida por $\mathcal{D}^n(\sigma) = \sigma$. Queremos mostrar que este par é δ-confiável. Temos

$$\begin{aligned} F\left(\mathcal{D}^n \circ \mathcal{C}^n, \rho^{\otimes n}\right) &= |\mathrm{Tr}(\rho^{\otimes n} P(n,\varepsilon))|^2 + \sum_i |\mathrm{Tr}(\rho^{\otimes n} A_i)| \\ &\geq |\mathrm{Tr}(\rho^{\otimes n} P(n,\varepsilon))|^2, \end{aligned}$$

mas pelo teorema do subespaço típico existe n_o tal que $\mathrm{Tr}(\rho^{\otimes n} P(n,\varepsilon)) < 1 - \frac{\delta}{2}$ para $n > n_o$. Segue que para $n > n_o$ o par é δ-confiável.
Para provar a recíproca suponha que $R < S(\rho)$. Sem perda de generalidade suponha que a operação de compressão mande $V^{\otimes n}$ em um subespaço de dimensão 2^{nR} com projetor correspondente $\Pi(n)$. Se o processo de compressão $\mathcal{C}_n$ é dado por operadores $\{C_{n,i}\}$ e o processo de descompressão $\mathcal{D}_n$ por $\{D_{n,i}\}$, temos

$$F\left(\mathcal{D}_n \circ \mathcal{C}_n, \rho^{\otimes n}\right) = \sum_{jk} |\mathrm{Tr}(D_{n,k} C_{n,j} \rho^{\otimes n})|^2.$$

Note que cada um dos operadores $C_{n,j}$ satisfaz a equação $C_{n,j} = \Pi(n) C_{n,j}$. Seja $\Pi^k(n)$ projetor sobre o subespaço obtido pela imagem de $\Pi(n)$ por $D_{n,k}$. Segue

$$\Pi^k(n) D_{n,k} \Pi(n) = D_{n,k} \Pi(n),$$

e assim $D_{n,k} C_{n,j} = D_{n,k}\Pi(n) C_{n,j} = \Pi^k(n) D_{n,k} \Pi(n) C_{n,j} = \Pi^k(n) D_{n,k} C_{n,j}$, de onde concluímos que

$$F\left(\mathcal{D}_n \circ \mathcal{C}_n, \rho^{\otimes n}\right) = \sum_{jk} |\mathrm{Tr}\left(D_{n,k} C_{n,j} \rho^{\otimes n} \Pi^k(n)\right)|^2.$$

Aplicando a desiguladade de Cauchy-Schwarz obtemos

$$F\left(\mathcal{D}_n \circ \mathcal{C}_n, \rho^{\otimes n}\right) \leq \sum_{jk} \mathrm{Tr}(D_{n,k} C_{n,j} \rho^{\otimes n} C_{n,j}^\dagger D_{n,k}^\dagger) \mathrm{Tr}(\Pi^k(n) \rho^{\otimes n}),$$

e o item 3 do teorema dos subespaços típicos nos diz que para qualquer $\delta > 0$ e n suficientemente grande, $\mathrm{Tr}\left(\Pi^k(n)\rho^{\otimes n}\right) \leq \delta$, unifirmemente em k. Assim temos

$$F\left(\mathcal{D}_n \circ \mathcal{C}_n, \rho^{\otimes n}\right) \leq \delta \sum_{jk} \mathrm{Tr}(D_{n,k} C_{n,j} \rho^{\otimes n} C_{n,j}^\dagger D_{n,k}^\dagger) = \delta,$$

que mostra que esse esquema não pode ser confíavel. □

Com este resultado encerramos este capítulo, mas fazemos duas ressalvas: uma que apenas tocamos superficialmente nos resultados da crescente área de informação quântica; outra que escolhemos apresentar este teorema, que é uma versão quântica do seu análogo clássico, mas o que torna a teoria interessante é que nem todas as analogias com o caso clássico são válidas. Também tocamos superficialmente nesta distinção no Exercício 3.25. De fato, este resultado é o caminho, por exemplo, para se ter um protocolo de criptografia quântica comprovadamente seguro[9].

Capítulo 4

Entropia e Mecânica Estatística

O objetivo deste capítulo é fazer uma breve introdução à Mecânica Estatística do Equilíbrio a volume finito (veja a Seção 4.1) com o intuito de provar o Teorema 4.2. Este teorema é o equivalente, no equilíbrio, do Teorema H de Boltzmann. O Teorema H de Boltzmann afirma que a entropia H de um gás de partículas, que está sob a ação de uma força externa atenuante e dependente do tempo, tende a crescer à medida que o tempo passa. Por outro lado, a Segunda Lei da Termodinâmica afirma que os sistemas físicos isolados tendem ao equilíbrio ao longo do tempo. Assim sendo, somos levados a concluir do teorema, juntamente com a Segunda Lei, que os estados de equilíbrio de um sistema físico são aqueles de maior entropia. No jargão popular, *os sistemas físicos fora do equilíbrio evoluem para o estado de maior entropia.*

O Teorema 4.2 é a manifestação deste comportamento, supondo de antemão que o sistema já esteja em equilíbrio termodinâmico. Em poucas palavras, o teorema afirma que o máximo da entropia, vista como função das probabilidades que descrevem o sistema, é atingido quando a probabilidade é aquela fornecida pelo regra de Gibbs (ou Boltzmann), veja a Definição 4.1.

Este capítulo é auto-contido. Contudo, caso o leitor queira se aprofundar no assunto, sugerimos a leitura das referências [3, 7, 8], de onde foi retirado o material para escrever estas notas.

4.1 Definições Preliminares

No que se segue, Λ representará um hipercubo em $\mathbb{Z}^d$, como definido a seguir: dado $N \in \mathbb{N}$, seja $\Lambda \equiv [-N, N]^d \cap \mathbb{Z}^d$. Contudo, deve-se ter em mente que Λ poderia ser um subconjunto arbitrário de $\mathbb{Z}^d$. A escolha do hipercubo é feita somente para facilitar a exposição. O conjunto Λ depende de N e é chamado de *rede*. Os seus elementos são chamados de *sítios*. Denotaremos por $|\Lambda|$ a *cardinalidade* de Λ, isto é, o número de sítios em Λ. Neste texto trabalharemos somente com *redes finitas* ($|\Lambda| < \infty$). As questões aqui tratadas podem ser reformuladas no *limite termodinâmico* (isto é, no limite $|\Lambda| \to \infty$), mas nós não faremos isto pois os pré-requisitos para entendê-las são bem mais avançados.

Exercício 4.1 *Para $N = 1$ e $d = 1, 2$ e 3, faça um desenho da rede Λ e calcule $|\Lambda|$. Deduza que, em geral, $|\Lambda| = (2N + 1)^d$.*

A cada sítio $i \in \Lambda$ nós associamos uma variável aleatória $\sigma_i \in \{-1, 1\}$, chamada de *spin*, com distribuição uniforme. Um ponto $\sigma \in \Omega \equiv \{-1, 1\}^{|\Lambda|}$ é uma *configuração* de spins e Ω é o *espaço das configurações*.

Exercício 4.2 *Para $N = 1 = d$, determine todas as possíveis configurações de spin. Para valores arbitrários de N e d, mostre que a cardinalidade de Ω é igual a $2^{|\Lambda|}$.*

Exercício 4.3 *Se $S \in \mathbb{Z}^+$ é dado e se σ é uma variável aleatória assumindo valores em $\{-S, -S + 1, \cdots, -1, 0, +1, \cdots, S - 1, S\}$, determine Ω e $|\Omega|$.*

A cada configuração $\sigma \in \Omega$ nós associamos uma *energia* (ou *Hamiltoniano*) $E_\Lambda(\sigma)$. Então $E_\Lambda(\sigma)$ é uma função de Ω em $\mathbb{R}$.

Exercício 4.4 *Se $N = 1 = d$, calcule os possíveis valores de $E_\Lambda(\sigma)$, onde*

$$E_\Lambda(\sigma) \equiv -(\sigma_0\sigma_1 + \sigma_0\sigma_{-1}).$$

A expressão da energia como função de σ é parte importante dos chamados *modelos* em mecânica estatística. Para a compreensão das próximas seções deste capítulo, não

é importante entender modelos específicos, embora sejam importantes exemplos (nesse contexto você será convidado a resolver o Exercício 4.7).

Denotaremos por β ao inverso de $K\,T$, onde $K > 0$ é a *constante de Boltzmann* e T é a *temperatura*. Veremos β como um parâmetro, sendo introduzido na teoria via o *fator de Gibbs* da configuração σ

$$e^{-\beta E_\Lambda(\sigma)}.$$

Sobre Ω nós definimos uma *medida de probabilidade* $\mu_{\Lambda,\beta}(\cdot)$, também chamada de *medida de Gibbs a volume finito* (associada à energia E_Λ e com parâmetro β)

$$\mu_{\Lambda,\beta}(\sigma) = \frac{e^{-\beta E_\Lambda(\sigma)}}{Z_\Lambda(\beta)}, \tag{4.1}$$

onde $Z_\Lambda(\beta)$ é um fator de normalização tal que $\mu_{\Lambda,\beta}(\Omega) = 1$. $Z_\Lambda(\beta)$ é a *função partição* e ela só depende dos parâmetros da teoria. A construção da medida 4.1 é o que chamamos de *Ensemble Canônico*. Se tivéssemos olhado para configurações de energia constante, teríamos o *Ensemble Microcanônico*.

Exercício 4.5 *Da definição de $\mu_{\Lambda,\beta}(\cdot)$, conclua que*

$$Z_\Lambda(\beta) = \sum_{\sigma\in\Omega} e^{-\beta E_\Lambda(\sigma)}.$$

Se f é uma função das configurações $\sigma \in \Omega$, então sua *esperança*, ou seu *valor esperado*, ou ainda seu *valor médio* (com respeito a $\mu_{\Lambda,\beta}(\cdot)$) é

$$\langle f\rangle_\Lambda = \sum_{\sigma\in\Omega} f(\sigma)\,\mu_{\Lambda,\beta}(\sigma). \tag{4.2}$$

É também comum representar o valor esperado de f por $\mu_{\Lambda,\beta}(f)$.

Definição 4.1 (Hamiltoniano de Ising) *O Hamiltoniano*

$$E_\Lambda(\sigma) \equiv -\sum_{\langle i,j\rangle} \sigma_i\sigma_j - h\sum_{i\in\Lambda}\sigma_i, \tag{4.3}$$

onde a primeira soma é feita sobre pares de sítios vizinhos dentro do volume Λ, onde $\sigma = \pm 1$ e onde $h \in \mathbb{R}$, é chamado de "Hamiltoniano de Ising com campo magnético externo h e com condições de contorno livres".

Observação O termo *condições de contorno livres* se refere ao fato de que as variáveis de spin dentro do volume Λ não interagem com as variáveis de spin externas a Λ (equivalentemente, no Hamiltoniano 4.3 não há parcelas envolvendo um produto $\sigma_i\sigma_j$, com $i \in \Lambda$ e $j \in \Lambda^c$). Outras condições de contorno são possíveis, além da livre. Por exemplo, se adicionarmos o termo $-\sum_i \sigma_i$, onde $i \in \partial^-\Lambda \equiv$ fronteira interna de Λ, ao Hamiltoniano 4.3, teremos o Hamiltoniano de Ising com *condições de contorno positivas*, significando que o spin em $i \in \partial^-\Lambda$ interage com um spin em Λ^c cujo valor é $+1$. De maneira similar, teremos *condições de contorno negativas*. Também poderiamos ter *condições de contorno periódicas*.

Exercício 4.6 *Supondo que $\sigma_i = \pm 1 = \sigma_j$, mostre que*

1. *$(\sigma_i\sigma_j)^n$ é igual a 1 se n é par e igual a $\sigma_i\sigma_j$ se n é ímpar;*
2. *$e^{\beta\sigma_i\sigma_j} = \cosh\beta + (\sinh\beta)\sigma_i\sigma_j$;*
3. *$\sum_{\sigma_i} 1 = 2$, $\sum_{\sigma_i} \sigma_i = 0$.*

Exercício 4.7 (Modelo de Ising Unidimensional) *Considere o Hamiltoniano (4.3), com $h = 0$ e em uma dimensão espacial. Começando a somar das variáveis de spin mais externas para as mais internas e usando o Exercício 4.6, compute explicitamente a função partição para o Modelo de Ising e mostre que*

$$Z(\beta) = 2(2\cosh\beta)^{2N}.$$

Mostre também que, para qualquer Λ e para qualquer par $i, j \in \Lambda$, vale a identidade

$$\langle\sigma_i\sigma_j\rangle_\Lambda = (\tanh\beta)^{|i-j|}.$$

4.2 O Princípio Variacional

A observação feita após a Definição 4.1 nos induz a pensar que, dado um sistema físico, podemos associar ao mesmo várias medidas de probabilidade: a cada condição de contorno imposta no sistema temos um Hamiltoniano associado que, por (4.1), gera uma medida de Gibbs. É claro que existem outras maneiras de se gerar medidas de probabilidade associadas ao sistema em questão. Por exemplo, poderíamos supor que as variáveis de

spin sejam independentes e uniformemente distribuída, ou ainda assumir uma distribuição de Bernoulli, com probabilidade $p > 1/2$ de $\sigma = +1$ ocorrer.

Portanto, dado um sistema físico, vamos denotar por $\mathcal{M}$ o conjunto de todas as medidas de probabilidade associadas ao mesmo e definidas sobre o espaço das configurações Ω. Fixado $f : \Omega \to \mathbb{R}$, podemos definir um *funcional* $\ell_f(\cdot) : \mathcal{M} \to \mathbb{R}$ como $\ell_f(\mu) \equiv \mu(f)$. Podemos agora reencontrar o conceito central destas notas:

Definição 4.2 *A* entropia de Boltzmann para medidas de probabilidade *é dada pelo funcional* $H : \mathcal{M} \longrightarrow \mathbb{R}$ *tal que*

$$\mu \quad \longmapsto \quad -\sum_{\omega\in\Omega} \mu(\omega) \ln \mu(\omega) \equiv H(\mu). \tag{4.4}$$

Na definição acima, usamos o logaritmo neperiano ao invés do logaritmo na base 2 (que foi usado nos capítulos anteriores) por causa das suas boas propriedades quando tomada a derivada. De qualquer maneira, se tivéssemos usado a base 2 então a única mudança no que se segue seria carregar uma constante em várias expressões.

Exercício 4.8 (Entropia de uma Medida de Gibbs) *Usando a Definição 4.2, mostre que a entropia de uma medida de Gibbs de energia $E_\Lambda(\sigma)$, a energia média (com respeito a essa medida de Gibbs) e a função partição estão associadas através da fórmula*

$$H(\mu_{\Lambda,\beta}) = \mu_{\Lambda,\beta}(\beta E_\Lambda) + \ln Z_\Lambda(\beta). \tag{4.5}$$

Fixados Λ, β e E_Λ, a igualdade (4.5) nos induz a definir o seguinte funcional sobre $\mathcal{M}$

$$p_{\Lambda,\beta}(\mu) = H(\mu) - \beta\mu(E_\Lambda). \tag{4.6}$$

A equação (4.5) nos diz que, se $\mu = \mu_{\Lambda,\beta}$ então $p_{\Lambda,\beta}(\mu_{\Lambda,\beta}) = \ln Z_\Lambda(\beta)$. Por causa disto o funcional $p_{\Lambda,\beta}(\cdot)$ é chamado de *pressão a volume finito e a temperatura K/β*. Contudo, é importante ressaltar que, do ponto de vista da Mecânica Estatística Clássica, a pressão a volume finito é definida como $|\Lambda|^{-1} \ln Z_\Lambda(\beta)$.

O Teorema H de Boltzmann afirma que os processos termodinâmicos não-estacionários ocorrem de maneira a maximizar a entropia. A entropia será máxima quando o equilíbrio termodinâmico for atingido. Provaremos, abaixo, uma versão deste fato para a Mecânica Estatística do Equilíbrio (veja o Teorema 4.2). Contudo, antes disto, vamos provar que a medida de Gibbs a volume Λ, com energia $E_\Lambda(\sigma)$ e à temperatura K/β é a única medida de probabilidade em $\mathcal{M}$ que maximiza a pressão a volume finito (4.6).

Teorema 4.1 (O Princípio Variacional) *Fixados Λ, β e E_Λ, seja $\mu_{\Lambda,\beta}(\cdot)$ a medida de Gibbs a volume finito Λ associada à energia E_Λ e com parâmetro β. Então*

$$\sup_{\mu\in\mathcal{M}} p_{\Lambda,\beta}(\mu) = H(\mu_{\Lambda,\beta}) - \beta\mu_{\Lambda,\beta}(E_\Lambda), \tag{4.7}$$

e $\mu_{\Lambda,\beta}(\cdot)$ é a única medida em $\mathcal{M}$ onde ocorre o supremo.

Prova: Para qualquer $\mu \in \mathcal{M}$, a pressão $p_{\Lambda,\beta}(\mu)$ pode ser reescrita como

$$p_{\Lambda,\beta}(\mu) = H(\mu) - \beta\mu(E_\Lambda) = -\sum_\sigma \mu(\sigma)\ln\mu(\sigma) - \beta\sum_\sigma E_{\Lambda,\beta}(\sigma)\mu(\sigma) =$$

$$-\sum_\sigma \mu(\sigma)\left[\ln\mu(\sigma) + \beta E_{\Lambda,\beta}(\sigma)\right] = \sum_\sigma \mu(\sigma)\ln\left(\frac{e^{-\beta E_{\Lambda,\beta}(\sigma)}}{\mu(\sigma)}\right).$$

Usando agora que $\ln x$ é uma função estritamente côncava e que vale uma desigualdade de Jensen para funções côncavas (veja Exercício 2.7), concluímos que

$$p_{\Lambda,\beta}(\mu) \leq \ln\left(\sum_\sigma \mu(\sigma)\left(\frac{e^{-\beta E_{\Lambda,\beta}(\sigma)}}{\mu(\sigma)}\right)\right) = \ln Z_\Lambda(\beta) = p_{\Lambda,\beta}(\mu_{\Lambda,\beta}).$$

Portanto, o máximo da pressão a volume finito é atingido em $\mu = \mu_{\Lambda,\beta} \in \mathcal{M}$. Com isto, provamos a existência do ponto de máximo. Para a unicidade, observe que, ainda pelo Exercício 2.7, a desigualdade acima torna-se uma igualdade se e somente se a razão

$$\frac{e^{-\beta E_{\Lambda,\beta}(\sigma)}}{\mu(\sigma)}$$

é constante para todo $\sigma \in \Omega$. Como μ é uma medida de probabilidade, é fácil concluir que, neste caso, a constante de proporcionalidade tem que ser necessariamente a função partição. Dito de outra forma, a desigualdade acima torna-se uma igualdade se e somente se

$$\frac{e^{-\beta E_{\Lambda,\beta}(\sigma)}}{Z_\Lambda(\beta)} = \mu(\sigma) \;\; \forall \;\; \sigma \in \Omega.$$

□

Exercício 4.9 *Prove o Teorema 4.1 usando multiplicadores de Lagrange. Para isto, considere a pressão a volume finito como função das $2^{|\Lambda|}$ variáveis $\mu(\sigma)$, $\sigma \in \Omega$. Determine o seu ponto de máximo, estando ela sujeita à restrição $\sum_\sigma \mu(\sigma) = 1$.*

Definição 4.3 (Estado de Equilíbrio) *Uma medida de probabilidade $\mu \in \mathcal{M}$ é um estado de equilíbrio associado à energia E_Λ se o supremo de (4.6) é atingido em μ.*

Observação: Segue então do Teorema 4.1 e da Definição 4.3 que toda medida de Gibbs a volume finito, associada à energia E_Λ, é um estado de equilíbrio.

Em Termodinâmica definem-se os *potenciais termodinâmicos* ou *energias livres*. Toda a descrição termodinâmica de um sistema físico é dada a partir do conhecimento do potencial termodinâmico adequado. Dentre eles destacamos a *energia livre de Gibbs*

$$F_{\Lambda,T}(\mu) = \mu(E_\Lambda) - KTH(\mu). \tag{4.8}$$

É claro que podemos reformular o Princípio Variacional 4.1 em termos desta energia livre: a medida de Gibbs $\mu_{\Lambda,\beta}$ minimiza a energia livre de Gibbs, sendo a única medida minimizante, isto é

$$F_{\Lambda,T}(\mu) \geq F_{\Lambda,T}(\mu_{\Lambda,\beta}) = -KT \ln Z_\Lambda(\beta).$$

Usando a definição da função partição a volume finito Λ e com energia $E_\Lambda(\sigma)$ (veja Exercício 4.5) e usando a regra da cadeia, podemos deduzir as seguintes relações para as derivadas da pressão a volume finito da medida de Gibbs $\mu_{\Lambda,\beta}$:

Exercício 4.10 *Mostre que, para qualquer valor de $\beta \in \mathbb{R}$, valem as indentidades:*

$$\frac{d}{d\beta} \ln Z_\Lambda(\beta) = -\mu_{\Lambda,\beta}(E_\Lambda), \quad \frac{d^2}{d\beta^2} \ln Z_\Lambda(\beta) = \mu_{\Lambda,\beta}(E_\Lambda^2) - [\mu_{\Lambda,\beta}(E_\Lambda)]^2. \tag{4.9}$$

Exercício 4.11 *Usando que $|\Omega| < \infty$, que $\mu_{\Lambda,\beta}(\Omega) = 1$ e a desigualdade de Cauchy-Schwartz em $\mathbb{R}^{|\Omega|}$, conclua que*

$$\mu_{\Lambda,\beta}(E_\Lambda^2) - [\mu_{\Lambda,\beta}(E_\Lambda)]^2 \geq 0.$$

Em particular, conclua que a pressão a volume finito de uma medida de Gibbs é uma função convexa (côncova) de β (de T).

Exercício 4.12 (Entropia-Energia) *Mostre que*

$$\frac{d}{dT} H(\mu_{\Lambda,\beta}) = \frac{1}{KT} \frac{d}{dT} \mu_{\Lambda,\beta}(E_\Lambda) \tag{4.10}$$

e conclua, novamente, que a pressão a volume finito é uma função côncova de T.

Observação: A identidade (4.10) é a forma mais clara de se exprimir uma relação infinitesimal muito conhecida em Termodinâmica:

$$dS = \frac{dQ}{T}.$$

Essa relação exprime o fato que a realização de trabalho vem sempre acompanhado de um aumento de entropia. O Exercício 4.12 mostra que esta relação pode ser obtida, de certa maneira, das definições dadas anteriormente.

O Teorema 4.1 tem o seguinte corolário

Teorema 4.2 (Entropia Maximal) *Seja E^* um número real entre o mínimo e o máximo da energia $E_\Lambda(\sigma)$, $\sigma \in \Omega$. Existe um único valor β^* de β tal que a medida de Gibbs a volume finito, associada à energia $E_\Lambda(\sigma)$ e com valor de pârametro β^*, satifaz à condição $\mu_{\Lambda,\beta^*}(E_\Lambda) = E^*$ e tal que $\mu_{\Lambda,\beta^*}(\cdot)$ maximiza a entropia em $\mathcal{M}$.*

Prova: O valor médio $\mu_{\Lambda,\beta}(E_\Lambda)$ é, claramente, uma função contínua de β pois, pelas relações (4.1) e (4.2), o valor médio é a razão de duas funções contínuas, sendo que o denomindor nunca se anula pois ele é uma soma finita de exponenciais. Pelos exercícios 4.10 e 4.11, $\mu_{\Lambda,\beta}(E_\Lambda)$ é uma função decrescente de β. Portanto, existem os limites $\lim_{\beta\to-\infty}\mu_{\Lambda,\beta}(E_\Lambda)$ e $\lim_{\beta\to+\infty}\mu_{\Lambda,\beta}(E_\Lambda)$. É claro que as desigualdades $\lim_{\beta\to-\infty}\mu_{\Lambda,\beta}(E_\Lambda) \leq \max_\sigma E_\Lambda(\sigma)$ e $\lim_{\beta\to+\infty}\mu_{\Lambda,\beta}(E_\Lambda) \geq \min_\sigma E_\Lambda(\sigma)$ são satisfeitas. Na verdade, vale sempre a igualdade pois, como tanto o numerador quanto o denominador são somas finitas envolvendo exponenciais de β, o limite da razão é determinado pelo termo exponencial dominante ($\exp[-\beta\max_\sigma E_\Lambda(\sigma)]$ se $\beta \to -\infty$ e $\exp[-\beta\min_\sigma E_\Lambda(\sigma)]$ se $\beta \to +\infty$).

Portanto, seja E^* um número real tal que $\min_\sigma E_\Lambda(\sigma) < E^* < \max_\sigma E_\Lambda(\sigma)$. Pela continuidade e monotonicidade de $\mu_{\Lambda,\beta}(E_\Lambda)$ como função de β, existe um valor β^* tal que $\mu_{\Lambda,\beta^*}(E_\Lambda) = E^*$. Agora, seja $\mu \in \mathcal{M}$ tal que $\mu(E_\Lambda) = E^*$. Então, pelo Princípio Variacional teremos que

$$H(\mu) - \beta^* E^* = H(\mu) - \beta^*\mu(E_\Lambda) \leq H(\mu_{\Lambda,\beta^*}) - \beta^*\mu_{\Lambda,\beta^*}(E_\Lambda) \leq H(\mu_{\Lambda,\beta^*}) - \beta^* E^*.$$

Conseqüentemente, $H(\mu) \leq H(\mu_{\Lambda,\beta^*})$ para qualquer $\mu \in \mathcal{M}$ satisfazendo $\mu(E_\Lambda) = E^*$.

□

Referências Bibliográficas

[1] W. Feller, Introdução à Teoria da Probabilidade, Editora Interciência, Rio de Janeiro (1978).

[2] B. James, Probabilidade: Um curso em nível intermediário, 2a ed., Projeto Euclides, Rio de Janeiro (1996).

[3] G. Keller, Equilibrium States in Ergodic Theory, London Mathematical Society Student Texts **42**, Cambridge University Press, Cambridge (1998).

[4] E. L. Lima, Curso de Análise, vol 1, 1a ed., Projeto Euclides, Rio de Janeiro (1976).

[5] M. A. Nielsen e I. L. Chuang, Quantum Computation and Quantum Information, Cambridge University Press, Cambridge (2000).

[6] A. N. Shiryaev, Probability, 2nd ed., Springer-Verlag, Berlin (1989).

[7] B. Simon, The Statistical Mechanics of Lattice Gases, Princeton University Press, Princeton (1993).

[8] C. Thompson, Mathematical Statistical Mechanics, The Macmillan Company, New York (1972).

[9] A.K. Ekert, "Quantum criptography based on Bell's theorem", Phys. Rev. Lett. **67**, 661 (1991).

[10] Quantum Cryptography Technology Experts Panel, A Quantum Information Science and Technology Roadmap, disponível em http://qist.lanl.gov (2004).

Índice Remissivo

Cadeia de Markov, 7
Combinação convexa, 15
Condições de contorno, 50
Configuração de spins, 48
Conjunto convexo, 15
Covariância, 9

Decomposição
  a valor singular, 37
  de Schmidt, 36
  ortonormal, 29
  polar, 37
Desigualdade
  de Chebyshev, 9
  de Jensen, 16
  de Klein, 31
Distribuição
  conjunta, 18
  de probabilidades, 6
  uniforme, 6
  binomial, 7

Energia, 48
Ensemble
  canônico, 49
  microcanônico, 49
Entropia
  binária, 14
  condicional, 19
  condicional quântica, 38
  conjunta, 18
  de Boltzmann, 51
  de Shannon, 13
  de uma medida de Gibbs, 51
  de von Neumann, 30, 31
  maximal, 15, 54
  relativa, 16
  relativa quântica, 30
Espaço
  amostral, 2
  das configurações, 48
  de estados, 26
  ε-típicos, 42
Esperança, 8
Estados, 26
  de equilíbrio, 53
  maximamente misturados, 30
  puros, 27
  ε-típicos, 42
Eventos, 2
  coletivamente independentes, 5
  independentes, 4

Fórmula de Bayes, 4
Fator de Gibbs, 49
Fidelidade, 44
Função
  côncava, 16
  convexa, 16
  indicadora, 6
  partição, 49, 51

Hamiltoniano, 48
  de Ising, 49
Hipercubo, 48

Informação mútua, 20, 38

Lei dos Grandes Números, 10

Medição local, 35
Medição projetiva, 28
Medida
    de Gibbs, 49
    de probabilidade, 2, 49
Modelo de Ising, 50

Observáveis, 28
Operador
    de medição, 27
    densidade, 26
    positivo, 25

Partição, 4
Ponto extremal, 15
Pressão, 51
Probabilidade condicional, 3
Processo
    compressão-descompressão, 43
    confiável, 41, 43
    de medição, 27
    de preparação, 27
    quântico, 44
Produto
    tensorial de espaços, 34
    tensorial de operadores, 34
Purificação, 37

Qubit, 26, 29, 36, 41

Sítios, 48
Seqüências ε-típicas, 21
Sistemas
    bipartites, 34
    compressão-descompressão, 23
    confiável, 23
    multipartites, 36
Subespaços ε-típicos, 42

Teorema
    concavidade da entropia, 16
    da multiplicação, 3
    da positividade da entropia relativa, 17
    da probabilidade total, 4
    das seqüências típicas, 21
    de Schumacher, 44
    de Shannon, 23
    do princípio variacional, 52
    do subespaço típico, 42
    entropia conjunta, 38
    H de Boltzmann, 47
    medições projetivas, 33
    subaditividade da entropia, 18, 38
Traço parcial, 35

Variáveis de spin, 48
Variáveis aleatórias, 5
    independentes, 6
    de Bernoulli, 6
Variância, 9
Vetor distribuição de probabilidade, 2

σ-álgebra, 1