LEONARDO ARAÚJO

TEORIA DA INFORMAÇÃO E CODIFICAÇÃO

Introdução

O avanço das tecnologias de comunicação trouxe a necessidade de buscar a transmissão de informação de forma eficiente. Informação, um conceito abstrato, que, no contexto da comunicação e da tecnologia, advém da organização e interpretação de dados, fornecendo uma visão sobre tendências ou padrões, possibilitando a interpretação e emergência de significado. Dados em si, são apenas um conjunto de símbolos, geralmente organizados como uma sequência. Assim, determinada informação pode ser representada na forma de dados de maneiras distintas. A representação na forma de dados pode ser mais ou menos propícia a um determinado meio, sob o qual a informação será transmitida ou armazenada. Uma mesma informação pode ainda ser representada de forma mais concisa ou prolixa.

Se voltarmos na história, em meados do século XIX, Samuel Morse e Alfred Vail propuseram o código Morse, como uma forma econômica de se comunicar através das redes telegráficas. A proposta consistia em usar um código de pontos (tom curto), traços (tom longo) e espaços de separação entre eles (silêncio) para tornar a comunicação de uma mensagem mais eficiente, ou seja, utilizando menos pulsos e, assim, reduzindo o tempo necessário para enviar uma mensagem. O código Morse, criado originalmente para o inglês, representa os símbolos mais frequentes através de sequências curtas e os símbolos infrequentes através de sequências longas, o que proporciona a redução do comprimento esperado da mensagem transmitida.

Também no século XIX, foi criada a escrita noturna por Charles Barbier. Esta forma de escrita foi posteriormente adaptada por Louis Braille para criar um sistema mais simples e acessível para deficientes visuais. Cada célula de 6 pontos é capaz de representar letras (ou sequências de letras) na forma de combinações binárias. O código Braille, inicialmente proposto para o francês, foi mais tarde adaptado para outras línguas, bem como para matemática e música, dentre outras áreas. A adaptação da forma de escrita e leitura ao meio é fundamental e evidente na forma escrita tátil, sendo preponderante para garantir a eficácia da comunicação. É essencial que o leitor seja capaz de decodificar facilmente a informação ali representada. Para tanto, a

distinção de diferentes símbolos é facilitada pela clareza e simplicidade do sistema. A eficiência e a economicidade da representação são evidentes na capacidade de transmitir informações de forma compacta, reduzindo o espaço necessário.

Usualmente, quando falamos da representação de informação, pensamos na escrita como uma forma de expressá-la. Entretanto, devemos nos atentar ao fato de que certas informações podem utilizar outros formatos representacionais. Por exemplo, a cotação de uma moeda é representada por uma sequência de números; um sinal eletrocardiográfico é medido pela diferença de potencial; uma imagem digital é constituída por uma matriz de pixels; e informações sobre produtos, endereços ou dados de pagamento podem ser armazenados em códigos de barras e códigos QR. Embora a natureza representacional dessas informações, ao serem geradas, não seja expressa por meio de um alfabeto convencional, elas podem ser transcodificadas para serem representadas através de um alfabeto padrão, como o Base64[1], que é utilizado para codificar, por exemplo, anexos de e-mails.

[1] Base64 é um esquema de codificação que transforma dados binários em uma representação textual utilizando um conjunto de 64 caracteres, que inclui letras maiúsculas (A-Z), letras minúsculas (a-z), dígitos (0-9) e os símbolos '+' e '/'. Cada grupo de três bytes de dados binários é convertido em quatro caracteres ASCII.

A ideia de representação binária é muito antiga, possivelmente anterior aos estudos de de Thomas Harriot e Gottfried Leibniz, nos séculos XVI e XVII. No entanto, Hartley (1928) foi um dos primeiros a quantificar a informação introduzindo o conceito de 'bit' como a unidade básica. Os trabalhos de Hartley ocorreram no contexto de rápida expansão das tecnologias de comunicação, como o telégrafo e o rádio, onde havia uma crescente necessidade de medir e otimizar a transmissão de dados.

O trabalho de Shannon (1948), que começou a ser desenvolvido durante a Segunda Guerra Mundial, permaneceu sigiloso devido à sua aplicação em comunicações militares. Embora suas ideias tenham sido formuladas na década de 1940, o artigo seminal *A Mathematical Theory of Communication* foi publicado apenas em 1948. Motivado pela necessidade de entender como a informação poderia ser codificada e transmitida de forma eficiente, Shannon desenvolveu uma teoria matemática que abordava questões práticas da comunicação. Ele introduziu conceitos fundamentais, como a entropia, que mede a incerteza ou a quantidade de informação em uma mensagem, e a capacidade do canal, que determina a quantidade máxima de informação que pode ser transmitida sem erro. O trabalho de Shannon estabeleceu um novo campo de estudo que influenciou profundamente a computação, a teoria da comunicação e outras disciplinas. Seu trabalho é considerado o marco de surgimento da teoria da informação.

Na década de 1960, a teoria da codificação passou por avanços significativos que moldaram a forma como entendemos e aplicamos a comunicação digital. O trabalho de Hamming (1950) estabeleceu as bases para a detecção e correção de erros, permitindo que sistemas de

comunicação se tornassem mais robustos. Paralelamente, os códigos de Golay (1949) emergiram como uma solução eficaz para a correção de múltiplos erros, ampliando as possibilidades de transmissão confiável. A aplicação do teorema de Shannon sobre a capacidade do canal continuou a ser explorada, fornecendo uma compreensão crítica dos limites da comunicação eficiente. Além disso, os códigos de convolução começaram a ganhar destaque, oferecendo novas abordagens para melhorar a confiabilidade na transmissão de dados. O trabalho inovador de Gallager (1962), com os códigos de paridade de baixa densidade, introduziu uma nova classe de códigos que se mostraram extremamente eficazes na correção de erros, influenciando profundamente o desenvolvimento da teoria da codificação. Esses avanços não apenas solidificaram a teoria da codificação como um campo essencial da teoria da informação, mas também tiveram um impacto duradouro em diversas aplicações práticas na comunicação moderna.

A teoria da informação tornou-se central nas comunicações digitais, servindo como a base para o desenvolvimento de tecnologias que transformaram a forma como nos comunicamos. Com a ascensão da internet e das redes de comunicação, os princípios estabelecidos por Claude Shannon, como a quantificação da informação e a capacidade dos canais, tornaram-se indispensáveis para otimizar a transmissão de dados e garantir a integridade das comunicações. Além de seu papel fundamental nas telecomunicações, a teoria da informação é amplamente aplicada em diversas outras áreas. Na ecologia, por exemplo, é utilizada para analisar a diversidade de espécies e a complexidade dos ecossistemas, ajudando a entender as interações entre organismos. Na criptografia, os conceitos de entropia e codificação são essenciais para garantir a segurança das informações transmitidas. Na linguística, a teoria da informação auxilia na análise da estrutura e da semântica das línguas, permitindo a modelagem de padrões de comunicação. Outros campos, como a biologia, onde a informação genética é estudada, e a psicologia, que investiga a percepção e a cognição, também se beneficiam dos princípios da teoria da informação, demonstrando sua relevância e aplicabilidade em diferentes áreas.

Princípios Fundamentais

A comunicação envolve alguns componentes fundamentais: uma fonte, que gera a mensagem; um transmissor, que envia a mensagem através de algum meio; um canal de comunicação, que é o meio pelo qual a comunicação se estabelece, sendo suscetível a ruídos e interferências; um receptor, que recebe a mensagem; e, por fim, o destinatário da mensagem (veja o diagrama na Figura 1, adapatado de "*A Mathematical Theory of Communication*"Shannon 1948). No âmago do processo de comunicação reside um problema fundamental: reconstruir no receptor a exata mensagem pretendida pelo emissor. Independentemente de estarmos lidando com a comunicação falada, escrita ou por meio de sinais digitais, o objetivo é o mesmo. Embora cada sistema guarde suas nuances, a teoria da informação proposta por Shannon (1948) busca lidar com essa problemática sob uma perspectiva unificada. Para tanto, ele introduziu conceitos fundamentais como entropia e capacidade de canal. Neste livro, procuraremos apresentar uma visão geral da teoria, seus fundamentos e aspectos práticos.

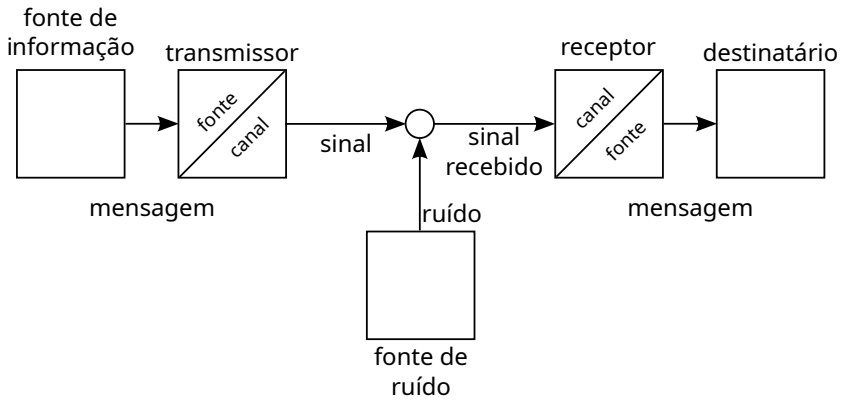

Figura 1: Diagrama genérico de um sistema de comunicações.

O diagrama ilustrado na Figura 1 apresenta a sistematização proposta por Shannon (1948), incluindo conceitos e aspectos da comunicação já utilizados à época, mas trazendo uma abordagem sistemática para o conceito de comunicação. Além disso, ele divide a tarefa do codificador e decodificador em duas partes distintas: a codifica-

ção/decodificação de fonte e a codificação/decodificação de canal. Essa separação permite abordar cada uma dessas partes do sistema de comunicação de forma independente, possibilitando, assim, projetar e otimizar cada uma delas separadamente.

Para lidar matermaticamente com a comunicação, um processo que envolve incerteza e variabilidade, e trazer então uma perspectiva probabilística na abordagem do problema, inicialmente, é essencial entender o conceito de variável aleatória (v.a.). Uma variável aleatória é uma variável que pode assumir diferentes valores dependendo de fatores aleatórios, ou seja, opera-se um mecanismo não determinístico que torna impossível prever seu valor. Ela representa a incerteza inerente a eventos ou processos aleatórios e pode ser utilizada para modelar diversos experimentos, como o lançamento de um dado, a previsão do tempo ou a sequência de caracteres em um texto. Na notação aqui adotada, as variáveis aleatórias serão representadas por letras maiúsculas, como, por exemplo, X. Uma variável aleatória discreta é aquela que pode assumir valores em um conjunto enumerável. As realizações de uma v.a. se dão de acordo com uma distribuição subjacente p, e assim dizemos que $X \sim p$. Para representar o conjunto de valores que a variável aleatória pode assumir, utilizaremos a letra em forma caligráfica $\mathcal{X}$, enquanto um valor específico que a variável pode assume será denotado por uma letra minúscula, como x. Assim, podemos descrever que a variável aleatória X assume o valor x através do evento $\{X = x\}$.

Dada uma v.a. X, em um alfabeto $\mathcal{X}$ de cardinalidade $N = |\mathcal{X}|$, onde $\mathcal{X} = \{a_1, \ldots, a_N\}$ e a_i, $i = 1, \ldots, N$, representa cada um dos possíveis valores que a v.a. pode assumir com probabilidade p_i, ou seja, $p_i = \Pr(X = a_i)$. A distribuição subjacente que rege a v.a. X é $p = \mathcal{P}_X = \{p_1, \ldots, p_N\}$, tal que $p_i \geq 0$ e $\sum_{i=1}^{N} p_i = 1$ ou, de forma equivalente, $\sum_{x \in \mathcal{X}} \Pr(X = x) = 1$.

Hartley (1928) introduziu a ideia de que a quantidade de informação pode ser medida em termos de possibilidades. Hartley propôs que a informação é diretamente proporcional ao logaritmo do número de estados possíveis de um sistema, estabelecendo assim uma base para a quantificação da informação. A medida de informação associada a uma variável aleatória X, com alfabeto de tamanho N, é expressa por

$$I(X) = \log_b L, \tag{1}$$

onde b é a base utilizada para medir tal informação. Para $b = 2$, a informação será medida em 'bits' (nome sugerido por J.W. Tukey).

O conceito proposto por Hartley sobre a quantificação da informação é diretamente consonante com a interpretação da capacidade representacional de uma unidade de memória. Em termos práticos, uma unidade de memória com n bits é capaz de representar 2^n se-

quências binárias distintas, o que implica que essa unidade possui n bits de informação. Quando consideramos duas unidades de memória, cada uma com n bits, a capacidade total se torna $2n$ bits, permitindo a representação de 2^{2n} sequências binárias distintas. Essa relação demonstra que, conforme a quantidade de bits de memória aumenta, a capacidade de representação de informações também cresce exponencialmente, alinhando-se perfeitamente com a proposta de Hartley de que a informação é medida em função do número de estados possíveis.

Um conceito fundamental que emerge na discussão sobre a teoria da informação é a redundância, ou ainda, a previsibilidade. A redundância refere-se à parte da informação que é repetitiva ou desnecessária para a compreensão, podendo ser eliminada sem perda de significado. Enquanto, sob uma primeira vista, pode ser vista como um desperdício de recursos, a redundância também desempenha um papel crucial na detecção e correção de erros, permitindo que a informação seja transmitida de forma mais confiável em ambientes ruidosos. Exemplo disso é a redundância presente na comunicação escrita. Veja como o trecho a seguir pode ser facilmente compreendido, mesmo com lacunas: "Nda se mudria, o regim, si era posívl, ms tmbém s muda d roupa sm trcar d pele."[2] A capacidade de compreender o trecho, apesar das palavras truncadas e da ausência de letras, demonstra que a linguagem é intrinsecamente projetada para lidar com a incerteza e a imperfeição no processo de transmissão. Se não houvesse redundância, qualquer perda de informação tornaria a decodificação da mensagem inviável, resultando em confusão e mal-entendidos.

[2] Trecho origial: "Nada se mudaria; o regime, sim, era possível, mas também se muda de roupa sem trocar de pele. Comércio é preciso. Os bancos são indispensáveis. No sábado, ou quando muito na segunda-feira, tudo voltaria ao que era na véspera, menos a constituição." (cap. LXIV de Esaú e Jacó, Machado de Assis).

Shannon, por outro lado, incorporou o conceito de redundância em sua análise da informação. Ao considerar eventos com probabilidades distintas, é importante notar que a incerteza associada a esses eventos não pode ser a mesma. Assim, a incerteza não depende apenas do tamanho do alfabeto, mas também das probabilidades associadas a cada evento. Para um evento E_k com probabilidade p_k, a quantidade de informação (medida em bits) associada a esse evento é expressa pela fórmula

$$I(E_k) = -\log_2(p_k). \tag{2}$$

Essa relação indica que eventos menos prováveis, ou seja, aqueles com probabilidades mais baixas, geram uma maior quantidade de informação, pois sua ocorrência é mais inesperada. Em contrapartida, eventos com alta probabilidade resultam em uma quantidade menor de informação, refletindo a redundância presente na comunicação. Essa compreensão da relação entre probabilidade e informação é essencial para o desenvolvimento do conceito de entropia, que serve como uma medida da incerteza, aleatoriedade e quantidade de informação associada a uma variável aleatória.

Entropia

Shannon (1948) propôs a entropia com forma de quantificar a incerteza associada a uma v.a., fornecendo assim uma medida da aleatoriedade. A entropia nos permite entender como a distribuição de probabilidades de uma variável aleatória influencia na informação associada a ela. A entropia de uma v.a. é então expressa como o valor esperado da informação própria associada aos eventos no alfabeto desta v.a., ou seja, $E_p[I(E_k)]$ e será expressa por $H(X)$.

Definição 1 (Entropia)**.**

$$H(X) \triangleq \mathrm{E}_p\left[-\log(p_k)\right] \tag{3a}$$

$$= -\sum_{k=1}^{|\mathcal{X}|} p_k \log p_k \tag{3b}$$

$$= -\sum_{x\in\mathcal{X}} p(x) \log p(x), \tag{3c}$$

Passamos a adotar a convenção $\log = \log_2$, uma vez que estaremos sempre medindo a informação em bits. Quando necessário utilizar uma base diferente, expressaremos explicitamente.

Para encontrar a formulação dada na Equação (3), Shannon (1948) propôs uma função de entropia H, que depende da distribuição de massa p. Dado um alfabeto de tamanho $N = \mathcal{X}$, e distribuição $p = \{p_1, \ldots, p_N\}$, Shannon estabeleceu três propriedades fundamentais que a função de entropia deveria satisfazer. A primeira propriedade exige que H seja contínua em relação às probabilidades p_i, o que significa que pequenas mudanças nas probabilidades não devem causar grandes saltos na medida de informação. A segunda propriedade afirma que, se todos os p_i são iguais, a entropia deve aumentar monotonicamente com o tamanho do alfabeto N. Isso reflete a intuição de que mais opções disponíveis para escolha devem resultar em maior incerteza e, portanto, maior entropia. Por fim, a terceira propriedade, conhecida como a propriedade da aditividade, estabelece que se uma escolha pode ser decomposta em uma sequência de escolhas independentes, a entropia total deve ser a soma ponderada das entropias individuais de cada escolha. Essa propriedade é crucial para a análise de sistemas complexos, onde a informação pode ser tratada em partes menores e combinadas para obter uma visão global.

A formulação matemática da entropia de Shannon, dada na Equação (3), emerge naturalmente ao considerar essas propriedades, sendo a função logarítmica utilizada devido à sua capacidade de transformar multiplicações em somas, o que é essencial para garantir a aditividade da medida de informação.

Interlúdio

Este interlúdio fornece uma pausa para apresentar a demonstração contida no Anexo 2 do artigo de Shannon (1948). O leitor, se desejar, pode ir direto à página 10, onde finaliza-se este entreato.

Define-se a $A(N)$ como o caso específico de H para uma distribuição uniforme com alfabeto de tamanho N:

$$A(N) = H\left(\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N}\right). \tag{4}$$

Deseja-se que uma escolha dentre s^M opções igualmente prováveis possa ser decomposta como uma sequência de M escolhas que se subdividem em s possibilidades igualmente prováveis.

Teremos então que

$$A(s^M) = MA(s). \tag{5}$$

Da mesma forma, para t e N, teremos $A(t^N) = NA(t)$. Podemos tomar N arbitrariamente grande e encontrar M que satisfaça

$$s^M \leq t^N \leq s^{(M+1)}. \tag{6}$$

Tomando o logaritmo[3] da expressão acima e dividindo por $N \log s$ todos os termos[4], teremos

$$\frac{M}{N} \leq \frac{\log t}{\log s} \leq \frac{M}{N} + \frac{1}{N}, \tag{7}$$

o que é equivalente a

$$\left|\frac{M}{N} - \frac{\log t}{\log s}\right| < \epsilon, \tag{8}$$

onde ϵ é arbitrariamente pequeno, já que N é arbitrariamente grande.

Usando agora a propriedade desejada de monotonicidade de $A(N)$, teremos

$$A(s^M) \leq A(t^N) \quad \leq A(s^{(M+1)}) \tag{9a}$$

$$M\,A(s) \leq N\,A(t) \leq (M+1)A(s) \tag{9b}$$

Dividindo a expressão acima por $NA(s)$, teremos

$$\frac{M}{N} \leq \frac{A(t)}{A(s)} \leq \frac{M}{N} + \frac{1}{N}, \tag{10}$$

ou, de forma equivalente,

$$\left|\frac{M}{N} - \frac{A(t)}{A(s)}\right| < \epsilon, \tag{11}$$

e assim, como as duas frações ($^{\log t}/_{\log s}$ e $^{A(t)}/_{A(s)}$) estão ϵ próximas de $^{M}/_{N}$, podemos concluir que

$$\left|\frac{A(t)}{A(s)} - \frac{\log t}{\log s}\right| < 2\epsilon. \tag{12}$$

Como ϵ é arbitrariamente pequeno, no limite teremos

$$\frac{A(t)}{A(s)} = \frac{\log t}{\log s} \tag{13a}$$

$$A(t) = \frac{A(s)}{\log s}\log t = K \log t, \tag{13b}$$

onde K deve ser positivo, de forma que $A(N)$ seja monótona crescente.

Suponha uma escolha com N possibilidades em que as probabilidades são comensuráveis, $p_i = {}^{N_i}/_{\sum N_i}$, onde N_i são inteiros. De forma equivalente,

[3] Logaritmo é uma função monótona crescente.

[4] $N \log s$ é positivo para $N \geq 0$ e $s \geq 1$.

uma escolha entre $\sum N_i$ opções pode ser expressa como uma escolha dentre N opções com probabilidades $p_1, \ldots, p_N$, e para uma i-ésima dada escolha, realizar uma nova escolha dentre N_i opções igualmente prováveis. Teremos então:

$$\overbrace{K \log\left(\sum N_i\right)}^{A(\sum N_i)} = H(p_1, \ldots, p_N) + \overbrace{K \log N_i}^{A(N_i)} \tag{14a}$$

$$K \underbrace{\left(\sum p_i\right)}_{=1} \log\left(\sum N_i\right) = H(p_1, \ldots, p_N) + K \underbrace{\left(\sum p_i\right)}_{=1} \log N_i. \tag{14b}$$

E assim,

$$H(p_1, \ldots, p_N) = K\left[\left(\sum p_i\right) \log\left(\sum N_i\right) - \left(\sum p_i\right) \log N_i\right] \tag{15a}$$

$$= -K \sum p_i \log \frac{N_i}{\sum N_i} = -K \sum p_i \log p_i. \tag{15b}$$

□

Tomemos primeiramente um exemplo simples de uma v.a. binária $X \in \mathcal{X} = \{0, 1\}$. Para simplificar, iremos denotar arbitrariamente por $\theta = Pr(X = 1)$ e $1 - \theta = Pr(X = 0)$. Usando agora a definição dada na Equação (3), teremos

$$H(X) = -\theta \log \theta - (1 - \theta) \log(1 - \theta) \tag{16}$$

e, para simplificar, iremos denotar esta grandeza por $H(\theta)$. Sempre que utilizarmos esta notação, como por exemplo em $H(1/4)$, $H(1/3)$ ou $H(\pi)$, estaremos nos referindo à entropia binária definida na Equação (16).

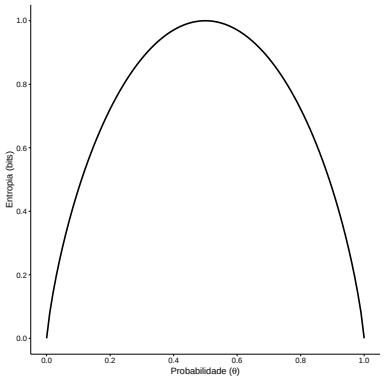

Figura 2: Entropia binária

A Figura 2 apresenta o gráfico da entropia binária dada pela Equação (16). Note como $H = 0$ quando $\theta = 1$ ou $\theta = 0$, ou seja, nos casos em que um dos eventos ocorre com probabilidade 1. Observe ainda como o máximo é alcançado em $\theta = 1/2$, quando ambos eventos são equiprováveis (máxima incerteza). A função observada é côncava, o que, mais à frente, mostraremos ser válido para toda função de entropia. A concavidade garante ainda que ao misturarmos (ponderarmos) distribuições diferentes não é possível obter uma entropia menor que a ponderação das entropias das partes.

Vamos considerar agora o exemplo de uma v.a. com distribuição uniforme $X \sim u$. Usando a definição na Equação (3), teremos que a entropia de X será dada por

$$H(X) = -\sum_{k=1}^{|\mathcal{X}|} \frac{1}{|\mathcal{X}|} \log \frac{1}{|\mathcal{X}|} \tag{17a}$$

$$= -\log 1/|\mathcal{X}| = \log |\mathcal{X}|. \tag{17b}$$

Como esperávamos, a entropia de uma distribuição uniforme é monotonicamente crescente com o tamanho do alfabeto. Intuitivamente,

como já mencionado, a entropia é máxima para uma distribuição uniforme, uma vez que é a situação de maior incerteza para um dado alfabeto.

Ao analisarmos a Equação (3), podemos nos questionar o que ocorre quando algum símbolo tem probabilidade zero, uma vez que aparecerá um termo $0 \log 0$ no somatório. Neste caso, iremos usar o valor limite $\lim_{x \to 0} x \log x = 0$. Para demonstrar (a menos de uma constate pela mudança de base do logaritmo), usaremos para tanto a regra de L'Hospital:

$$\lim_{x \to 0} x \ln x = \lim_{x \to 0} \frac{\ln x}{1/x} \tag{18a}$$

$$= \lim_{x \to 0} \frac{1/x}{-1/x^2} \tag{18b}$$

$$= \lim_{x \to 0} -x = 0. \tag{18c}$$

No próximo exemplo, vamos buscar aproximar a entropia de textos escritos. Sabemos que a distribuição das palavras em textos segue uma lei de potência conhecida como Lei de Zipf Zipf 1935; Zipf 1949; Ferrer i Cancho e Solé 2001; Araújo, Cristófaro-Silva e Yehia 2013. A Lei de Zipf é uma observação empírica que também é observada em diversos outros naturais. Ao analisar um corpus de texto, Zipf constatou que a frequência de uma palavra é inversamente proporcional à sua posição em um ranking de frequência. Essa relação pode ser expressa matematicamente como

$$p_k(s, N) = C\, k^{-s}, \tag{19}$$

onde p_k é a probabilidade de ocorrência da k-ésima palavra mais frequente, C é uma constante de normalização ($C^{-1} = \sum_{n=1}^{N} n^{-s}$), k é o *rank*, s a constante que caracteriza a distribuição e N o tamanho do léxico. A Figura 3 apresenta o gráfico de Zipf (frequência/probabilidade vs. *rank*) para a obra completa de William Shakespeare, dividida em 44 textos, conforme organizados no Projeto Gutenberg[5]. Cada uma das série de pontos na Figura 3 representa um destes 44 textos. Note como o comportamento observado entre eles é bem similar.

[5] https://www.gutenberg.org/

Usando a Equação (19) em conjunto com a Equação (3), obtemos a formulação da entropia para uma distribuição de Zipf:

$$H(X) = s\, C \sum_{k=1}^{N} \frac{\log k}{k^s} - \log C. \tag{20}$$

Após estimar a entropia de cada um dos textos de Shakespeare, podemos observar no histograma da Figura 4 como a concentração dos valores encontrados está por acima de 9 bits. Como observaremos mais

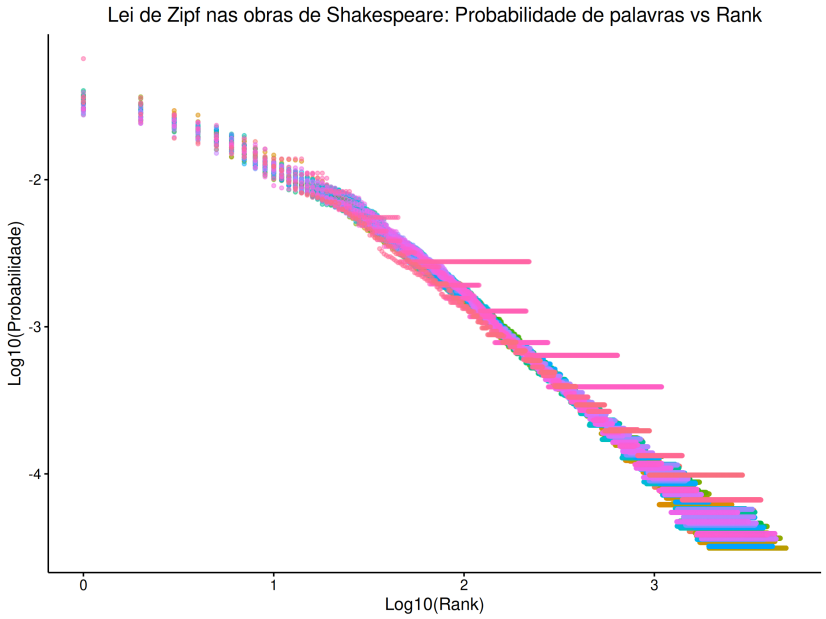

Figura 3: Gráfico de Zipf para 44 textos de William Shakespeare no Projeto Gutenberg.

à frente, conhecer o valor da entropia é importante para sabermos qual é o limite representacional para uma fonte (qual é o limite da compressão). Além disso, linguistas utilizam a entropia para analisar a complexidade de uma língua, redundância e eficiência do sistema de escrita.

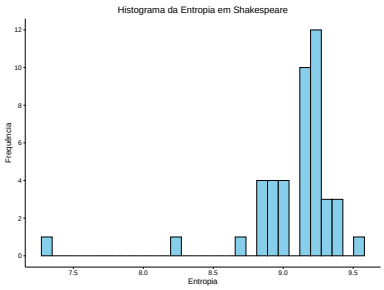

Figura 4: Histograma da entropia nos 44 textos de William Shakespeare.

Propriedades da Entropia

A entropia possui várias propriedades fundamentais que iremos descrever a seguir. Essas propriedades não apenas ajudam a entender o comportamento da entropia e medidas relacionadas, como também serão uteis em diversas demonstrações. Nesta seção, exploraremos as principais propriedades da entropia, incluindo a não negatividade, a concavidade, valor máximo, a continuidade, e a mudança de base.

Como dito anteriormente, usualmente a unidade escolhida para medir a informação é bit, e portanto utilizamos a base 2. Entretanto, a entropia pode ser dada também em outras bases: nats (base e), trits (base 3), hartley (base 10). Para obtermos a entropia em outras bases, basta utiliza a regra de mudança de base do logaritmo: $\log_a x = {}^{\log_b x}/_{\log_b a}$. Teremos assim $H_b(X) = (\log_b a) H_a(X)$. A transformação para a base e costuma ser útil nas demonstrações para evitar o fardo de se carregar uma constante durante todo o seu percurso.

Uma propriedade muito importante é a não negatividade, ou seja, $H(X) \geq 0$. A incerteza nunca[6] será negativa, o que faz sentido intui-

[6] Importante ressaltar que estamos aqui tratando de variáveis discretas. No contexto contínuo a entropia pode ser negativa e terá outra interpretação.

tivamente. Além disso, a não negatividade é uma propriedade importante para encontrarmos outras relações. Para verificá-la, basta notar que a entropia é a soma de termos da forma $-p \log p$, como $0 \leq p \leq 1$, teremos que cada termo da soma é maior ou igual a zero. Por conseguinte, a entropia seria não negativa.

Para verificar que a entropia é uma função contínua em p, devemos primeiramente definir $l(x)$ como a extensão contínua de $x \log x$, da seguinte forma:

$$l(x) = \begin{cases} x \log x & \text{se } x > 0, \\ 0 & \text{se } x = 0, \end{cases} \tag{21}$$

e assim, $H(X) = H(p)$ é definido como

$$H(p) = -\sum_{x \in \mathcal{X}} l(p_x). \tag{22}$$

Sendo a entropia definida como um somatória finito de funções contínuas, fica evidente que ela também será uma função contínua.

Antes de abordarmos uma próxima propriedade fundamental da entropia, vamos estabelecer uma desigualdade simples e importante em diversas demonstrações futuras. Ela é apresentada graficamente na Figura 5 e demonstrada formalmente no Teorema 1.

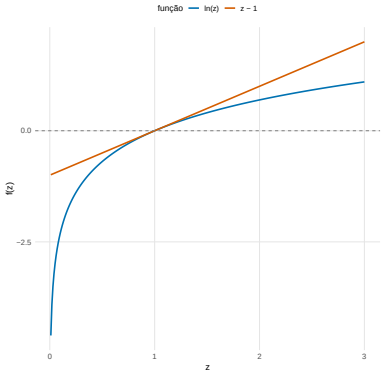

Figura 5: Demonstração gráfica da desigualdade $\ln z \leq z - 1$ (embora não seja uma demonstração formalmente válida, serve como intuição).

Lema 1 (Desigualdade Fundamental) *Para qualquer* $z > 0$,

$$\ln z \leq z - 1, \tag{23}$$

com igualdade se e somente se $z = 1$.

Demonstração. Sabemos que para $z = 1$ é verdadeiro, pois $0 = \ln 1 \leq 1 - 1 = 0$. Vamos demonstrar que $\ln z \leq z - 1$, para $z \geq 1$ por contradição. Suponha que existe $b > 1$ tal que $\ln z > z - 1$, para $z = b$. Vamos definir $f(z) = \ln z - z + 1$, logo $f(1) = 0$ (conforme visto acima) e $f(b) > 0$ (por hipótese). Pelo teorema do valor médio $\exists c$, $1 < c < b$, tal que

$$f'(c) = \frac{f(b) - f(1)}{b - 1} = \underbrace{\frac{\overbrace{f(b)}^{>0}}{b - 1}}_{>0} > 0. \tag{24}$$

Mas, $f'(z) = 1/z - 1$, e assim $f'(z) < 0$ para $z > 1$. Logo há uma contradição e nossa hipótese é falsa. Teremos assim $\ln z \leq z - 1$, para $z \geq 1$.

Para $z \in (0, 1)$, basta seguir os mesmos passos, escolhendo um ponto $z = b$, tal que $0 < b < 1$. Vamos encontrar um ponto c tal que $0 < b < c < 1$. E usar o teorema do valor médio para mostrar uma contradição na hipótese.

□

Outra propriedade importante que mencionamos é o limite máximo da entropia, a entropia da distribuição uniforme. Note que, mostrar que $H(X) \leq \log N$ é equivalente a mostrar que $H(X) - \log N \leq 0$. Para tanto, vejamos:

$$
\begin{aligned}
H(X) - \log N &= -\sum_x p(x) \log p(x) - \log N \overbrace{\sum_x p(x)}^{=1} && \text{(25a)}\\
&= -\sum_x p(x) \log p(x) - \sum_x p(x) \log N && \text{(25b)}\\
&= -\sum_x p(x) \log p(x)\, N && \text{(25c)}\\
&= \log_2 e \sum_x p(x) \ln \frac{1}{p(x)N} && \text{(25d)}\\
&\leq \log_2 e \sum_x p(x) \left[\frac{1}{p(x)N} - 1\right] && \text{(25e)}\\
&= \log_2 e \left[\underbrace{\sum_{x\in\mathcal{X}} \frac{1}{N}}_{=N\frac{1}{N}=1} - \underbrace{\sum_x p(x)}_{=1}\right] = 0. && \text{(25f)}
\end{aligned}
$$

□

Na demonstração anterior, utilizamos a Equação (23), com $z = {}^{1}\!/_{p(x)\,N}$. A igualdade $\ln z = z - 1$ se dará no ponto $z = 1$, isto é, quando ${}^{1}\!/_{p(x)\,N} = 1$, ou seja, quando $p(x) = {}^{1}\!/_{N}$, e teremos assim uma distribuição uniforme.

A concavidade da entropia é uma propriedade importante. Garante que a entropia terá um único máximo global. A maximização da entropia é um princípio fundamental em várias áreas como estatística, para a inferência de distribuições de probabilidade, e aprendizado de máquina, para garantir que os algoritmos de otimização convirjam e para garantir a estabilidade dos modelos. Esta propriedade será demonstrada mais à frente.

Entropia Conjunta e Entropia Condicional

Quando lidamos com incerteza em situações que envolvem múltiplas variáveis aleatórias, entropia conjunta e a entropia condicional são conceitos que surgem da extensão das definições vistas anteriormente. Um par de varáveis aleatórias (X, Y) pode ser visto como uma variável aleatória vetorial.

A entropia conjunta de duas variáveis aleatórias X e Y, denotada como $H(X, Y)$, mede a incerteza total associada ao par de variáveis. É definida como:

Definição 2 (Entropia Conjunta)**.**

$$H(X,Y) \triangleq -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \tag{26a}$$

$$= \mathrm{E}_{p(x,y)} \log \frac{1}{p(X,Y)} \tag{26b}$$

$$= \mathrm{E} \log \frac{1}{p(X,Y)}, \tag{26c}$$

onde, em 26c, para simplificar[7], omitimos a distribuição em respeito a qual o valor esperado é tomado.

[7] Tais simplificações pode aparecer ao longo do texto. Porém as adotaremos apenas em casos que não gerem ambiguidades.

Generalizando para um vetor de variáveis aleatórias $X_{1:N} = (X_1, X_2, \ldots, X_N)$:

$$H(X_{1:N}) = H(X_1, X_2, \ldots, X_N) \tag{27a}$$

$$= \sum_{x_1,x_2,\ldots,x_N} p(x_1,\ldots,x_N)\log\frac{1}{p(x_1,\ldots,x_N)} \tag{27b}$$

$$= \mathrm{E}\log\frac{1}{p(X_1,\ldots,X_N)}. \tag{27c}$$

Dada a definição, vamos analisar um exemplo prático para analisar a interdependência entre caracteres consecutivos em uma língua natural. Para calcular a entropia conjunta de dois caracteres em sequência, devemos primeiramente calcular a probabilidade conjunta. A Figura 6 apresenta o resultado encontrado para a obra completa de Shakespeare. Note como os maiores valores correspondem às sequências 'th', 'he', 'er', 'an', dentre outras (devido às palavras de alta frequência de ocorrência na língua, como "the", "in", "her", "there", etc.

Ao calcular a entropia conjunta encontramos $H(X,Y) = 7.8482$ bits, o que é menor do que o dobro da entropia de um único caractere, $2 \times H(X) = 8.3798$ bits. O fato da entropia conjunta de dois caracteres consecutivos ser menor do que o dobro da entropia de um único caractere indica que os caracteres não são independentes: a ocorrência de um caractere influencia a probabilidade do próximo, reduzindo a incerteza total. Em uma língua natural, essa interdependência é esperada devido a padrões linguísticos.

Essa interdependência sugere que a incerteza associada ao segundo caractere (Y), dado o conhecimento do primeiro caractere (X), é menor do que a incerteza do segundo (Y) quando considerado isoladamente. Em outras palavras, o conhecimento de X fornece informação que reduz a incerteza sobre Y, como ocorre em pares comuns como 'th', onde 't' aumenta a probabilidade de 'h'. Essa incerteza reduzida é formalmente chamada de entropia condicional, denotada por $H(Y|X)$, e é definida como

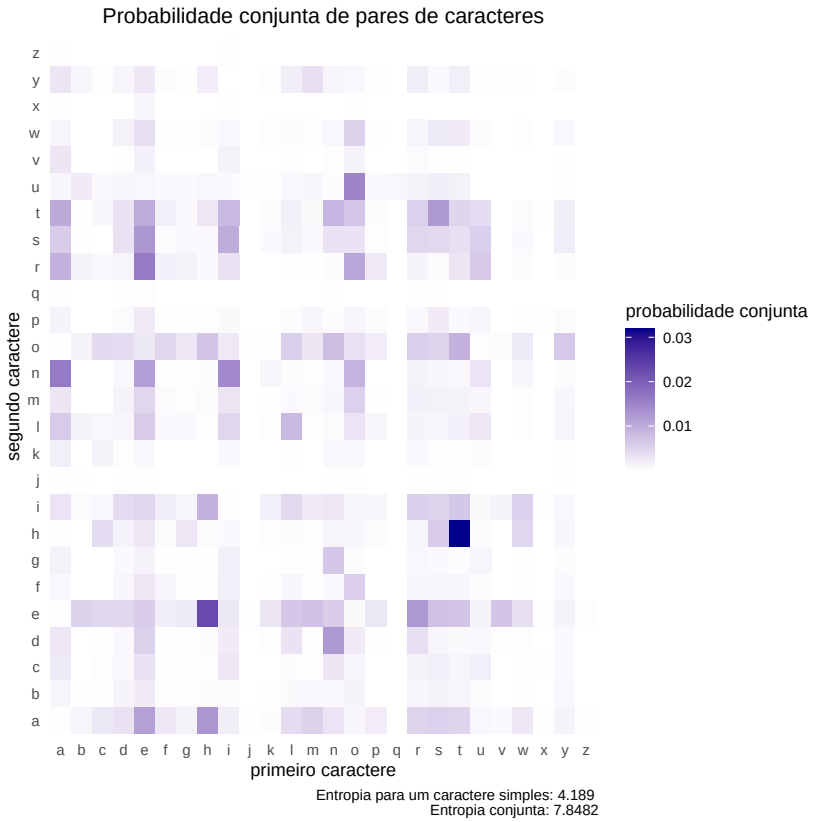

Figura 6: Probabilidade conjunta do primeiro e segundo caractere analisando os 44 textos de William Shakespeare no Projeto Gutenberg. As entropias simples e conjunta são apresentadas, evidenciando que a entropia conjunta é maior, porém, analisando a entropia por caractere, a entropia conjunta abarca dependências entre caracteres consecutivos, reduzindo o efeito de incerteza por caractere.

Definição 3 (Entropia Condicional)**.**

$$H(Y|X) \triangleq \sum_x p(x)H(Y|X=x) \tag{28a}$$

$$= -\sum_x p(x)\sum_y p(y|x)\log p(y|x) \tag{28b}$$

$$= -\sum_{x,y} p(x,y)\log p(y|x) \tag{28c}$$

$$= \mathrm{E}_{p(x,y)} \log \frac{1}{p(Y|X)}. \tag{28d}$$

A definição dada pela Equação (28a) nos mostra como a entropia condicional de Y, dado X, é o valor esperado da entropia de Y condicionada aos eventos x. Observamos anteriormente (e será demonstrado posteriormente) que $H(Y|X) \leq H(Y)$, pois conhecer X reduz a incerteza sobre Y. Entretanto, devemos ressaltar que nada podemos dizer com relação a $H(Y|X=x)$, podendo este ser menor, igual ou maior que $H(Y)$. Para um caso específico, condicionar pode aumentar a incerteza, embora, na média, a entropia condicional seja sempre menor ou igual a entropia sem ser condicionada. Um exemplo interessante, adaptado de Cover e Thomas (2006), é o contexto de um julgamento: uma evidência única evidência pode aumentar a incerteza sobre um determinado caso, o que pode levar a um julgamento enviesado; entretanto, conhecer um conjunto de evidência nos leva esperar uma incerteza menor, eliminando o viés no julgamento.

Note que, condicionando $H(Y|X)$ a Z, usando a Equação (28a) podemos escrever

$$H(Y|X,Z) = \sum_z p(z)H(Y|X,Z=z), \tag{29}$$

onde

$$H(Y|X,Z=z) = -\sum_{x,y} p(x,y|z)\log p(y|x,z). \tag{30}$$

Observando as definições de entropia conjunta e entropia condicional, surge, naturalmente, a constatação de que a entropia de um par de variáveis aleatória é a entropia de uma somada à entropia condicional da outra. Esta observação é formalizada pelo teorema da regra da cadeia.

Teorema 2 (Regra da Cadeia)

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{31}$$

Demonstração. Como $p(x,y) = p(x)p(y|x)$ (teorema de Bayes), e como o logaritmo de um produto é a soma dos logaritmos dos fatores, teremos

$$-\log p(x,y) = -\log p(x) - \log p(y|x) \tag{32}$$

onde multiplicamos por (-1) ambos os lados. Calculando o valor esperado de ambos os lados, obtemos o resultado desejado.

□

Exemplo 1. Considere duas variáveis aleatórias X e Y, com $\mathcal{X} = \{x_1, x_2, x_3\}$ e $Y = \{y_1, y_2, y_3\}$. A distribuição conjunta $p(x, y)$ é dada:

$p(x,y)$	y_1	y_2	y_3
x_1	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$
x_2	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
x_3	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{4}$

A entropia conjunta é dada por

$$\begin{aligned} H(X,Y) &= -2\frac{1}{4}\log\frac{1}{4} - 2\frac{1}{8}\log\frac{1}{8} - 3\frac{1}{16}\log\frac{1}{16} - 2\frac{1}{32}\log\frac{1}{32} \\ &= 1 + \frac{6}{8} + \frac{3}{4} + \frac{10}{32} = 2.8125 \text{ bits.} \end{aligned}$$

Para calcular $H(X)$ devemos primeiramente calcular a marginal

$$p(x) = \sum_y p(x,y) = \left\{\frac{7}{16}, \frac{7}{32}, \frac{11}{32}\right\}.$$

e assim podemos calcular a entropia de X

$$\begin{aligned} H(X) &= -\frac{7}{16}\log\frac{7}{16} - \frac{7}{32}\log\frac{7}{32} - \frac{11}{32}\log\frac{11}{32} \\ &= \frac{7}{16}(4 - \log 7) + \frac{7}{32}(5 - \log 7) + \frac{11}{32}(5 - \log 11) \\ &= \frac{146}{32} - \frac{21}{32}\log 7 + \frac{55}{32} - \frac{11}{32}\log 11 = 1.5310 \text{ bits.} \end{aligned}$$

Para calcular agora a entropia condicional $H(Y|Y)$, devemos calcular a entropia condicional $p(y|x)$

$p(y\|x)$	y_1	y_2	y_3
x_1	$\frac{4}{7}$	$\frac{4}{7}$	$\frac{2}{11}$
x_2	$\frac{2}{7}$	$\frac{2}{7}$	$\frac{1}{11}$
x_3	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{8}{11}$

Dada as duas distribuições $p(x, y)$ e $p(y|x)$, podemos agora utilizar a Equação (28c) para calcular $H(Y|X)$,

$$\begin{aligned} H(Y|X) &= -\frac{1}{4}\log\frac{4}{7} - \frac{1}{8}\log\frac{4}{7} - \frac{1}{16}\log\frac{2}{11} - \frac{1}{8}\log\frac{2}{7} - \frac{1}{16}\log\frac{2}{7} - \frac{1}{32}\log\frac{1}{11} - \frac{1}{16}\log\frac{1}{7} - \frac{1}{32}\log\frac{1}{7} - \frac{1}{4}\log\frac{8}{11} \\ &= -\frac{3}{8}\log\frac{4}{7} - \frac{3}{16}\log\frac{2}{7} - \frac{3}{32}\log\frac{1}{7} - \frac{1}{16}\log\frac{2}{11} - \frac{1}{32}\log\frac{1}{11} - \frac{1}{4}\log\frac{8}{11} \\ &= \frac{3}{8}(\log 7 - 2) + \frac{3}{16}(\log 7 - 1) + \frac{3}{32}(\log 7) + \frac{1}{16}(\log 11 - 1) + \frac{1}{32}(\log 11) + \frac{1}{4}(\log 11 - 3) \\ &= \frac{21}{32}\log 7 + \frac{11}{32}\log 11 - \frac{56}{32} \\ &= 1.2815 \text{ bits.} \end{aligned}$$

Com os resultados anteriores podemos verificar a regra da cadeia:

$$H(X,Y) = H(X) + H(Y|X) \quad \Rightarrow \quad 2.8125 = 1.5308 + 1.2817 \text{ bits.}$$

Mesmo para um exemplo simples como este, é mais prático calcular numericamente, como apresentado no código Octave a seguir:

```
p_xy = [1/4, 1/8, 1/16; 1/8, 1/16, 1/32; 1/16, 1/32, 1/4];
H_XY = -sum(sum(p_xy .* log2(p_xy)));
p_y = sum(p_xy, 1);
p_y_given_x = p_xy./p_y; % p(y|x);
H_Y_given_X = -sum(sum(p_xy.*log2(p_y_given_x)));
```

Informação Mútua e Entropia Relativa

Definimos anteriormente a entropia como uma medida da incerteza sobre uma variável aleatória, e a entropia condicional como a incerteza condicionada ao conhecimento de outra viável aleatória. Estas definições nos levam a questionar o quanto uma variável aleatória revela sobre a outra, ou, equivalentemente, o quanto a incerteza sobre uma é reduzida ao se conhecer a outra. Esta medida é dada pela informação mútua, definida como

Definição 4 (Informação Mútua)**.**

$$I(X;Y) \triangleq H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{33}$$

onde as duas igualdades são válidas, por simetria. A primeira reflete a redução da incerteza sobre X quando Y é conhecido, e, de forma similar, a segunda reflete a redução da incerteza sobre Y quando X é conhecido.

Note que a notação $I(X;Y)$ usa ponto e vírgula em vez de vírgula para enfatizar que a informação mútua mede a relação ou dependência entre as variáveis X e Y, distinguindo-se de uma função conjunta como $H(X,Y)$, onde a vírgula separa argumentos de uma distribuição.

A partir da definição dada na Equação (33), podemos reescrevê-la como a seguir:

$$\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) && \text{(34a)}\\
&= \mathrm{E}_{p(x)} \log \frac{1}{p(x)} - \mathrm{E}_{p(x,y)} \log \frac{1}{p(x|y)} && \text{(34b)}\\
&= \mathrm{E}_{p(x,y)} \log \frac{p(x|y)}{p(x)} && \text{(34c)}\\
&= \mathrm{E}_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)} && \text{(34d)}\\
&= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. && \text{(34e)}
\end{aligned}$$

A informação mútua, dada pela Equação (34e), pode ser vista como a divergência de Kullback–Leibler entre a distribuição conjunta e o produto das marginais, ou seja,

$$I(X;Y) = D(p(x,y)||p(x)p(y)). \tag{35}$$

Vejamos então a definição de divergência de Kullback–Leibler.

Definição 5 (Divergência de Kullback–Leibler)**.** A divergência de Kullback–Leibler (KL) entre duas distribuições p e q, em um alfabeto comum $\mathcal{X}$, é definida como

$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{36a}$$

$$= \mathrm{E}_p \log \frac{p(X)}{q(X)}. \tag{36b}$$

Note que, em geral, a divergência de KL não é simétrica, ou seja, $D(p||q) \neq D(q||p)$.

Na definição de acima, adotamos a convenção de que o somatório se dá sobre o suporte da distribuição p, S_p.

A divergência de KL é conhecida também na literatura como entropia relativa. Por não ser simétrica, não pode ser considerada como uma métrica ou distância, e ainda não satisfaz a desigualdade triangular.

Como observado anteriormente, seja $\mu_1(x,y) = p(x,y)$ (distribuição conjunta) e $\mu_2(x,y) = p(x)p(y)$ (produto das marginais), com $p(x) = \sum_y p(x,y)$ e $p(y) = \sum_x p(x,y)$, então

$$D(\mu_1||\mu_2) = \sum_{x,y} \mu_1(x,y) \log \frac{\mu_1(x,y)}{\mu_2(x,y)} \tag{37a}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = I(X;Y). \tag{37b}$$

A informação mútua é a distância entre a distribuição conjunta em X e Y e o produto das distribuições marginais em X e Y. Se as variáveis aleatórias são independentes, teremos $p(x,y) = p(x)p(y)$ e por conseguinte a divergência será nula, a informação mútua entre X e Y será zero. A informação mútua é o erro em se assumir independência entre as variáveis aleatórias.

O produto das distribuições marginais $p(x)p(y)$ é uma projeção da distribuição conjunta $p(x,y)$ sobre o conjunto dos produtos de distribuições independentes, minimizando a divergência entre a distribuição conjunta e o produto das marginais. Ou seja,

$$p(x)p(y) = \underset{p'(x,y) \backslash p'(x,y) = p'(x)p'(y)}{\operatorname{argmin}} D(p(x,y)||p'(x,y)). \tag{38}$$

Podemos agora considerar uma aplicação prática em problemas de estimação paramétrica: minimizar a divergência KL entre uma distribuição subjacente p_θ e uma distribuição empírica $\hat{p}$ revela-se equivalente a maximizar o logaritmo da verossimilhança dos dados sob o modelo $\hat{p}$. A divergência KL quantifica o custo ao usar $\hat{p}$ para aproximar p_θ, e otimizar esse custo nos leva diretamente a ajustar os parâmetros θ para melhor descrever os dados observados.

Seja $x_1, \ldots, x_N \in \mathcal{X}$, N observações i.i.d.[8] de uma variável aleatória X. A distribuição empírica será dada por

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n), \tag{39}$$

onde δ é a função de Dirac.

[8] Uma coleção de variáveis aleatórias é independente e identicamente distribuída (i.i.d.) se todas possuírem uma mesma distribuição e forem mutuamente independentes.

Seja p_θ uma distribuição em $\mathcal{X}$ parametrizada por θ. Maximizar a verossimilhança de $p_\theta(x)$ é equivalente a minimizar a divergência de KL $D_{\mathrm{KL}}(\hat{p} \parallel p_\theta)$.

$$\begin{aligned}
D_{\mathrm{KL}}(\hat{p} \parallel p_\theta) &= \sum_{x \in \mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} && \text{(40a)} \\
&= -H(\hat{p}) - \sum_{x \in \mathcal{X}} \hat{p}(x) \log p_\theta(x) && \text{(40b)} \\
&= -H(\hat{p}) - \frac{1}{N} \sum_{x \in \mathcal{X}} \sum_{n=1}^{N} \delta(x - x_n) \log p_\theta(x) && \text{(40c)} \\
&= -H(\hat{p}) - \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n) && \text{(40d)} \\
&= -H(\hat{p}) - \ell(p_\theta(x)), && \text{(40e)}
\end{aligned}$$

onde $\ell(\cdot)$ é a função log-verossimilhança.

A estimativa de máxima verossimilhança de θ a partir das N observações é dada por

$$\begin{aligned}
\hat{\theta}_n &= \operatorname*{argmax}_{\theta \in \Theta} \prod_{n=1}^{N} p_\theta(x_n) && \text{(41a)} \\
&= \operatorname*{argmax}_{\theta \in \Theta} \sum_{n=1}^{N} \log p_\theta(x_n) && \text{(41b)} \\
&= \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{n=1}^{N} -\log p_\theta(x_n). && \text{(41c)}
\end{aligned}$$

Desta forma, podemos constatar que a distribuição que minimiza a divergência de KL para a distribuição empírica é aquela que maximiza a verossimilhança (ou logaritmo desta). Pela lei forte dos grandes números[9], $\frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_n) \xrightarrow{q.c.} \mathrm{E}[\log p_\theta(X)]$.

[9] A Lei Forte dos Grandes Números afirma que, dada um sequência infinita $X_1, X_2, X_3, \ldots$ de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.) com média finita $\mathrm{E}[X_i] = \mu$, a média amostral $\overline{X}_n = {}^1\!/_n \sum_{i=1}^{n} X_i$ converge quase certamente para μ, ou seja, $\overline{X}_n \xrightarrow{q.c.} \mu$. .

Ao ajustar os parâmetros θ para minimizar divergência KL, estamos buscando o modelo $\hat{p}$ que melhor representa p. A não negatividade da

divergência, garante que o custo informacional nunca será negativo e que a aproximação ótima ocorre quando as distribuições coincidem. Graças à convexidade da função logarítmica, sabemos que esse ótimo é bem definido e único."

Vejamos então uma propriedade essencial da divergência: a não negatividade.

Teorema 3 (Não negatividade da Divergência) *Para duas distribuições p e q em um alfabeto comum $\mathcal{X}$,*

$$D(p||q) \geq 0 \tag{42}$$

com igualdade se e somente se $p = q$.

Demonstração. Se $q(x) = 0$ para algum $x \in S_p$, então $D(p||q) = \infty$ e o teorema é verdadeiro (caso trivial). Vamos assumir então que $q(x) > 0 \forall x \in S_p$. Teremos então

$$\begin{aligned} D(p||q) &= \log e \sum_{x \in S_p} p(x) \ln \frac{p(x)}{q(x)} && \text{(43a)} \\ &\geq \log e \sum_{x \in S_p} p(x) \left(1 - \frac{q(x)}{p(x)}\right) && \text{(43b)} \\ &= \log e \left[\sum_{x \in S_p} p(x) - \sum_{x \in S_p} q(x)\right] && \text{(43c)} \\ &\geq \log e(1 - 1) && \text{(43d)} \\ &= 0, && \text{(43e)} \end{aligned}$$

onde em 43b utilizamos a Equação (23) substituindo z por $1/z$ (o que fornece $\ln z \geq 1 - 1/z$) e para obter 43d utilizamos o fato de que

$$\sum_{x \in S_p} q(x) \leq 1, \tag{44}$$

uma vez que o somatório é tomado no suporte de p, podendo assim excluir alguns pontos não nulos de q.

□

O fato da divergência ser não negativa, nos leva como consequência em termos também a não negatividade da informação mútua, uma vez que esta é um caso específico de divergência.

Lema 4 (Não negatividade da Informação Mútua) *A informação mútua entre duas variáveis aleatórias X e Y é tal que*

$$I(X;Y) \geq 0, \tag{45}$$

com igualdade se e somente se $X \perp\!\!\!\perp Y$

Demonstração. A informação mútua é a divergência entre a distribuição conjunta e o produto das marginais, logo, aplicando o Teorema 3, obtemos

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0. \tag{46}$$

Quando $X \perp\!\!\!\perp Y$, teremos $p(x,y) = p(x)p(y)$ e assim a divergência será nula e também a informação mútua.

□

A divergência também pode ser utilizada para demonstrar o limite máximo da entropia, de forma alternativa àquela apresentada na Equação (25).

Teorema 5 (Limite Máximo da Entropia) $H(X) \leq \log|\mathcal{X}|$, *onde* $|\mathcal{X}|$ *denota a cardinalidade do alfabeto* $\mathcal{X}$*, com igualdade se e somente se* X *possuir distribuição uniforme.*

Demonstração. Seja $u(x) = 1/|\mathcal{X}|$ a função probabilidade de massa uniforme em $\mathcal{X}$, e seja $p(x)$ a função probabilidade de massa para X. Então

$$\begin{aligned} D(p||u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= \sum_{x\in\mathcal{X}} p(x)\log p(x) + \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} \\ &= -H(X) + \log|\mathcal{X}|\sum_{x\in\mathcal{X}} p(x) \\ &= \log|\mathcal{X}| - H(X). \end{aligned} \tag{47}$$

Como a entropia relativa é não negativa, $D(p||u) \geq 0$, teremos

$$D(p||u) = \log|\mathcal{X}| - H(X) \geq 0, \tag{48}$$

e assim

$$H(X) \leq \log|\mathcal{X}| \tag{49}$$

□

Observação 1 *Para uma variável aleatória em um alfabeto D-ário, ou seja,* $|\mathcal{X}| = D$*, a entropia na base* D *será*

$$H_D(X) \leq 1. \tag{50}$$

Observação 2 *O Teorema 5 estabelece um valor máximo para a entropia desde que o alfabeto seja finito. Para o caso de alfabeto de tamanho infinito, a entropia pode ser finita ou não.*

Vejamos dois exemplos:

Exemplo 2. Seja X uma variável aleatória com distribuição dada por

$$Pr(X = i) = 2^{-i}, \quad i = 1, 2, \ldots. \tag{51}$$

Neste caso, teremos

$$H(X) = \sum_{i=1}^{\infty} i2^{-i} = 2. \tag{52}$$

Exemplo 3 (Yeung 2002)**.** Considere uma variável aleatória X com valores em pares de inteiros

$$\left\{(i, j) : 1 \leq i \leq \infty \text{ e } 1 \leq j \leq \frac{2^{2^i}}{2^i}\right\} \tag{53}$$

tal que

$$Pr(X = (i, j)) = 2^{-2^i}, \tag{54}$$

para todo i e j. A entropia será dada por

$$H(X) = -\sum_{i=1}^{\infty} \sum_{j=1}^{2^{2^i}/2^i} 2^{-2^i} \log 2^{-2^i} = \sum_{i=1}^{\infty} 1 \tag{55}$$

que não converge.

Vejamos ainda algumas outras observações que podemos fazer sobre a informação mútua. Primeiramente, pela Definição 4, é fácil notar que a informação mútua $I(X;Y)$ é simétrica em X e Y, ou seja, $I(X;Y) = I(Y;X)$.

Em seguida, vamos analisar o que ocorre quando tomas a informação mútua entre uma variável aleatória X e elas mesma, ou seja, $I(X;X)$.

Proposição 1 (A informação mútua de uma variável aleatória e ela mesma é a entropia) *A informação mútua de uma variável aleatória X com ela mesma, a auto-informação de X, é igual à entropia de X.*

Demonstração.

$$\begin{aligned} I(X;X) &= \mathrm{E} \log \frac{p(X)}{p(X)^2} && \text{(56a)} \\ &= -\mathrm{E} \log p(X) && \text{(56b)} \\ &= H(X). && \text{(56c)} \end{aligned}$$

□

Podemos descrever uma nova relação, ao observar a Definição 4 ($I(X;Y) = H(X) - H(X|Y)$) e o Teorema 2 ($H(X|Y) = H(X,Y) - H(Y)$)

Proposição 2

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{57}$$

Todas essas relações observadas entre entropia, entropia conjunta e informação mútua podem ser facilmente representadas através do diagrama de Venn apresentado na Figura 7. Esta correspondência entre o diagrama de Venn e as medidas de informação de Shannon não são mera coincidência.

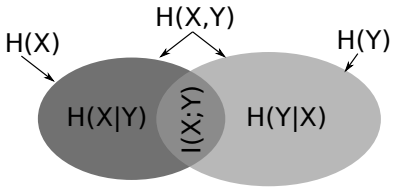

Figura 7: Diagrama de Venn ilustrando as relações entre entropias, entropias condicionais e informação mútua, facilitando a visualização das grandezas informacionais.

De forma semelhante que vimos anteriormente, quando foi introduzido o conceito de entropia condicional, podemos agora estender o conceito de informação mútua entre duas variáveis aleatórias X e Y dado o conhecimento de uma terceira variável Z, definindo assim a informação mútua condicional.

Definição 6 (Informação Mútua Condicional)**.** Dadas as variáveis aleatórias X, Y e Z, a informação mútua entre X e Y, condicionada ao conhecimento de Z é dada por

$$\begin{aligned}
I(X;Y|Z) &\triangleq \sum_z p(z) I(X;Y|Z=z) && \text{(58a)}\\
&= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x,y|Z=z)}{p(x|Z=z)p(y|Z=z)} && \text{(58b)}\\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} && \text{(58c)}\\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x|y,z)p(y|z)}{p(x|z)p(y|z)} && \text{(58d)}\\
&= \mathrm{E}_{p(x,y,z)} \left[\log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y,z)} \right] && \text{(58e)}\\
&= H(X|Z) - H(X|Y,Z). && \text{(58f)}
\end{aligned}$$

Generalização da Regra da Cadeia

A generalização é fundamental para analisar sistemas com múltiplas variáveis ou séries temporais. Uma sequência de variáveis aleatórias $X_1, X_2, \ldots, X_N = X_{1:N}$ pode representar estados ou observações ao longo do tempo. A regra da cadeira, vista anteriormente no Teorema 2, pode ser generalizada para uma sequência de N variáveis aleatórias $X_{1:N}$.

Proposição 3 (Regra da Cadeia Generalizada da Entropia) *Dada uma sequência de N variáveis aleatórias $X_{1:N}$, a entropia conjunta pode ser dada por*

$$H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}). \tag{59}$$

Demonstração. A demonstração é feita por indução. Sabemos que para $N = 2$ é verdadeiro, como visto no Teorema 2. Vamos supor que para $N = M$ é verdadeiro e mostrar que para $N = M + 1$ é verdadeiro.

$$H(X_1, X_2, \ldots, X_M, X_{M+1}) = H(X_1, X_2, \ldots, X_M) + H(X_{M+1}|X_1, X_2, \ldots, X_M) \tag{60a}$$

$$= \sum_{i=1}^{M} H(X_i|X_1, \ldots, X_{i-1}) + H(X_{M+1}|X_1, X_2, \ldots, X_M) \tag{60b}$$

$$= \sum_{i=1}^{M+1} H(X_i|X_1, \ldots, X_{i-1}), \tag{60c}$$

onde em 60a utilizamos o Teorema 2, fazendo $X = (X_1, X_2, \ldots, X_M)$ e $Y = X_{M+1}$, e em 60b utilizamos a premissa de que a relação 59 é válida para $N = M$. Isto prova então que a relação 59 é válida para $N = M + 1$ e, como é sabidamente válida para $N = 2$, será válida para todo N.

□

A regra da cadeia vista no Teorema 2 pode ser estendida para o caso do condicionamento à uma terceira variável aleatória.

Proposição 4 (Regra da Cadeia Condicional) *Para duas variáveis aleatórias X e Y condicionadas a uma terceira, Z, a entropia conjunta condicional pode ser decomposta como:*

$$H(X, Y|Z) = H(X|Z) + H(Y|X, Z). \tag{61}$$

A incerteza total sobre (X, Y), dado Z, pode ser quebrada em duas partes: primeiro, a incerteza sobre X dado Z, e depois a incerteza adicional de Y dado tanto X quanto Z.

Demonstração. A demonstração segue os mesmos passos da demonstração do Teorema 2, bastando condicionar a Z para obtermos a adaptação para a regra da cadeia condicional.

□

Fazemos agora a generalização para uma sequência de variáveis aleatórias $X_{1:N}$ condicionada a uma variável aleatória Y.

Proposição 5 (Regra da Cadeia Condicional Generalizada)

$$H(X_1, X_2, \ldots, X_N|Y) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}, Y). \tag{62}$$

Demonstração.

$$H(X_1, X_2, \ldots, X_N|Y) = \sum_y H(X_1, X_2, \ldots, X_N|Y = y) \tag{63a}$$

$$= \sum_y p(y) \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}, Y = y) \tag{63b}$$

$$= \sum_{i=1}^{N} \sum_y p(y) H(X_i|X_1, \ldots, X_{i-1}, Y = y) \tag{63c}$$

$$= \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}, Y), \tag{63d}$$

onde utilizamos 28a em 63a, 29 em 63d, e 63b segue de 59.

□

Por fim, podemos aplicar as mesmas ideias para obter a regra da cadeia para informação mútua, que descreve como a informação compartilhada entre um conjunto de variáveis e outra variável pode ser decomposta em contribuições condicionais.

Proposição 6 (Regra da Cadeia da Informação Mútua)

$$I(X_1, X_2, \ldots, X_N; Y) = \sum_{i=1}^{N} I(X_i; Y|X_1, \ldots, X_{i-1}). \tag{64}$$

Demonstração.

$$I(X_1, X_2, \ldots, X_N; Y) = H(X_1, X_2, \ldots, X_N) - H(X_1, X_2, \ldots, X_N|Y) \tag{65a}$$

$$= \sum_{i=1}^{N} [H(X_i|X_1, \ldots, X_{i-1}) - H(X_i|X_1, \ldots, X_{i-1}, Y)] \tag{65b}$$

$$= \sum_{i=1}^{N} I(X_i; Y|X_1, \ldots, X_{i-1}), \tag{65c}$$

onde utilizamos as Proposições 3 e 5.

□

Desigualdade de Jensen

Nesta seção, apresentamos a desigualdade de Jensen, uma relação observada para funções convexas entre o valor esperado da função de uma variável aleatória e a função do valor esperado desta variável aleatória. Mais precisamente, $\mathrm{E}f(X) \geq f(\mathrm{E}X)$, para f uma função convexa e X uma variável aleatória.

A desigualdade de Jensen é uma ferramenta fundamental para demonstrar diversas propriedades essenciais. Entretanto, primeiramente devemos rever o teste da segunda derivada para convexidade.

Definição 7 (Função Convexa)**.** Dizemos que f é convexa em (a, b) se para todo $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \tag{66}$$

A Figura 8 ilustra uma função convexa no intervalo, evidenciando como a imagem de um ponto intermediário $\lambda x_1 + (1-\lambda)x_2$, entre x_1 e x_2, é menor ou igual ao correspondente na corda que passa por $(x_1, f(x_1))$ e $(x_2, f(x_2))$.

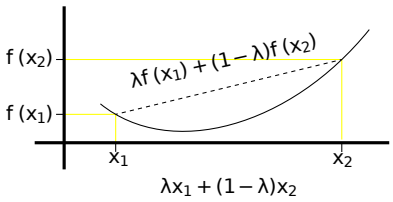

Figura 8: Ilustração de uma função convexa no intervalo.

Alguns exemplos de funções convexas são: $f(x) = x^2$; $f(x) = x^4$; $f(x) = e^x$; e $x \log x$, $x \geq 0$. Teremos que f é estritamente convexa se a igualdade for verdadeira apenas para $\lambda = 0$ ou $\lambda = 1$. Uma função f é dita côncava se $-f$ (f multiplicada por -1) for uma função convexa, ou seja, côncavo é o oposto de convexo.

Para determinar se uma função é convexa, podemos realizar o simples teste da segunda derivada.

Proposição 7 (Teste da Segunda Derivada para Convexidade) *Se uma função f possui derivada segunda não-negativa (positiva) em um intervalo, a função é convexa (estritamente convexa) no intervalo.*

Demonstração. A expansão de Taylor de uma função f em torno do ponto x_0 é dada por

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \tag{67}$$

onde $x^* \in (x_0, x)$. Por hipótese, $f''(x^*) \geq 0$, e desta forma, o último termo é não-negativo.

Seja $x_0 = \lambda x_1 + (1-\lambda)x_2$. Analisando em $x = x_1$, teremos

$$f(x_1) \geq f(x_0) + f'(x_0)(x_1 - \lambda x_1 - (1-\lambda)x_2) \tag{68a}$$
$$= f(x_0) + f'(x_0)((1-\lambda)(x_1 - x_2)) \tag{68b}$$

Da mesma forma, em $x = x_2$, teremos

$$f(x_2) \geq f(x_0) + f'(x_0)(x_2 - \lambda x_1 - (1-\lambda)x_2) \tag{69a}$$
$$= f(x_0) + f'(x_0)(\lambda(x_2 - x_1)) \tag{69b}$$

Somando λ 68b com $(1-\lambda)$ 69b, obtemos

$$\lambda f(x_1) + (1-\lambda)f(x_2) \geq \lambda f(x_0) + \lambda f'(x_0)((1-\lambda)(x_1 - x_2)) + (1-\lambda)f(x_0) + (1-\lambda)f'(x_0)(\lambda(x_2 - x_1)) \tag{70a}$$
$$\geq f(x_0) = f(\lambda x_1 + (1-\lambda)x_2) \tag{70b}$$

□

Podemos agora retornar à desigualdade de Jensen. Vemos na Figura 9 uma ilustração do que tal propriedade representar.

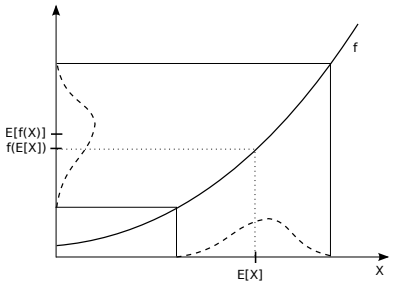

Figura 9: Ilustração da desigualdade de Jensen, fornecendo uma intuição para ela.

Teorema 6 (Desigualdade de Jensen) *Seja f uma função convexa e X uma variável aleatória, então*

$$\mathrm{E}f(X) = \sum_x p(x)f(x) \geq f(\mathrm{E}X) = f\left(\sum_x xp(x)\right) \tag{71}$$

Demonstração. Para uma distribuição de massa com apenas dois pontos

$$\mathrm{E}[f(X)] = p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) = f(\mathrm{E}X) \tag{72}$$

já que f é convexa e $p_1 + p_2 = 1$.

Para uma distribuição com mais de dois ponto, iremos fazer uma demonstração por indução.

Suponha que o teorema seja verdadeiro para uma distribuição com $k-1$ pontos de massa. Para uma distribuição com k pontos de massa podemos escrever cada $p'_i = p_i/(1-p_k)$ para $i = 1, 2, \ldots, k-1$.

Desta forma, teremos

$$\begin{aligned}
\mathrm{E}[f(X)] &= \sum_{i=1}^{k} p_i f(x_i) && \text{(73a)}\\
&= \sum_{i=1}^{k-1} (1-p_k) p'_i f(x_i) + p_k f(x_k) && \text{(73b)}\\
&= (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) + p_k f(x_k) && \text{(73c)}\\
&\geq (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) + p_k f(x_k) && \text{(73d)}\\
&\geq f\left((1-p_k) \sum_{i=1}^{k-1} p'_i x_i + p_k x_k\right) && \text{(73e)}\\
&= f\left(\sum_{i=1}^{k} p_i x_i\right) && \text{(73f)}
\end{aligned}$$

onde em 73c utilizamos a hipótese de indução, já que

$$\sum_{i=1}^{k-1} p'_i = \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} = \frac{1-p_k}{1-p_k} = 1. \tag{74}$$

e em 73d utilizamos a definição de convexidade.

Desta forma, sendo o teorema válido para uma distribuição de massa com $k-1$ pontos, também será verdadeiro para uma distribuição de massa com k pontos. Como mostramos que para $k = 2$ é verdadeiro, logo o teorema é verdadeiro para qualquer k.

□

A desigualdade de Jensen pode ser utilizada para demonstrar a não negatividade da divergência de KL.

Lema 7 (Não negatividade da Divergência de Kullback–Leibler)

$$D(p||q) \geq 0 \text{ } \textit{com igualdade se e somente se } p(x) = q(x) \forall x. \tag{75}$$

Demonstração. De forma equivalente, iremos mostrar que $-D(p||q) \leq 0$.

Seja $S_p = \{x : p(x) > 0\}$, o suporte de p, então

$$-D(p||q) = -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in S_p} p(x) \log \frac{p(x)}{q(x)} \tag{76a}$$

$$= \sum_{x \in S_p} p(x) \log \frac{q(x)}{p(x)} = \mathrm{E} \log \frac{q(X)}{p(X)} \tag{76b}$$

$$\leq \log \left(E \frac{q(X)}{p(X)} \right) = \log \left(\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} \right) \tag{76c}$$

$$= \log \left(\sum_{x \in S_p} q(x) \right) \leq \log \left(\sum_x q(x) \right) = \log 1 = 0 \tag{76d}$$

onde em 76b utilizamos a desigualdade de Jensen, Equação (71).

□

A desigualdade da soma dos logaritmos é derivada da desigualdade de Jensen e da convexidade da função logarítmica. Esta desigualdade também aparece em algumas demonstrações importantes.

Proposição 8 (Desigualdade da Soma de Logaritmos) *Dados* $(a_1, \ldots, a_n)$ *e* $(b_1, \ldots, b_n)$*, com* $a_i \geq 0$ *e* $b_i \geq 0$*, temos*

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i} \tag{77}$$

e teremos igualdade se e somente se $a_i/b_i = c$*, onde* c *é uma constante.*

Demonstração. Considere $f(t) = t \log t = t(\ln t)(\log e)$, que é estritamente convexa, pois $f''(t) = 1/t \log e > 0$, $\forall t > 0$.

Dada f convexa, a desigualdade de Jensen diz que

$$\sum_i \alpha_i f(t_i) \geq f \left(\sum_i \alpha_i t_i \right) \text{ com } \alpha_i \geq 0 \text{ e } \sum_i \alpha_i = 1. \tag{78}$$

$f(x) = x \log x$ é estritamente convexa para $x > 0$, já que $f''(x) = \frac{1}{x} \log e > 0$ para $x > 0$.

Vamos fazer $\alpha_i = {}^{b_i}/_{\sum_{j=1}^{n} b_j}$ e $t_i = {}^{a_i}/_{b_i}$, então obteremos

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \tag{79a}$$

$$\sum_i \left(\frac{b_i}{\sum_j b_j} f\left(\frac{a_i}{b_i}\right)\right) \geq f\left(\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\right) \tag{79b}$$

$$\frac{1}{\sum_j b_j}\sum_i \left(b_i \frac{a_i}{b_i}\log\frac{a_i}{b_i}\right) \geq \left(\sum_i \frac{a_i}{\sum_j b_j}\right)\log\sum_i \frac{a_i}{\sum_j b_j} \tag{79c}$$

$$\sum_i a_i \log\frac{a_i}{b_i} \geq \left(\sum_i a_i\right)\log\sum_i \frac{a_i}{\sum_j b_j} \tag{79d}$$

$$\sum_i a_i \log\frac{a_i}{b_i} \geq \sum_i a_i \log\frac{\sum_i a_i}{\sum_j b_j}. \tag{79e}$$

□

A desigualdade da soma de logaritmos pode ser utilizada para mostrar que $D(p||q) \geq 0$.

Proposição 9 (Não negatividade da Divergência) $D(p||q) \geq 0$

Demonstração.

$$D(p||q) = \sum_x p(x)\log\frac{p(x)}{q(x)} \tag{80a}$$

$$\geq \left(\sum_x p(x)\right)\log\frac{\sum_x p(x)}{\sum_x q(x)} \tag{80b}$$

$$= 1\log\frac{1}{1} = 0. \tag{80c}$$

□

A entropia relativa, ou divergência de Kullback-Leibler, possui uma propriedade importante: ela é convexa no par de distribuições. A convexidade no par considera a interpolação simultânea de ambas as distribuições e garante que a divergência KL se comporte de forma previsível ao combinar distribuições.

Teorema 8 (A Entropia Relativa é Convexa no Par) *Para dois pares de distribuições* (p_1, q_1) *e* (p_2, q_2),

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2), \tag{81}$$

para todo $0 \leq \lambda \leq 1$.

Demonstração. Usando a Definição 5, definição de divergência de KL, temos

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) = \sum_x (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \tag{82a}$$

$$\leq \sum_x \left(\lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \right) \tag{82b}$$

$$= \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \tag{82c}$$

onde, em 82a, podemos considerar que, dentro do somatório em x, temos um somatório com dois termos: $(a_1 + a_2) \log ({}^{(a_1 + a_2)}/_{(b_1 + b_2)})$. Utilizamos então a Equação (77) para o caso com apenas dois temos:

$$\left(\sum_{i=1}^{2} a_i \right) \log \frac{\sum_{i=1}^{2} a_i}{\sum_{i=1}^{2} b_i} \leq \sum_{i=1}^{2} a_i \log \frac{a_i}{b_i} = a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2}, \tag{83}$$

o que nos leva ao resultado em 82b. E usamos novamente a Definição 5 para obter 82c.

□

Teorema 9 (A Entropia Relativa é Convexa no Primeiro Argumento) *A divergência de Kullback-Leibler, $D(p||q)$, é convexa em relação a p, fixando-se q. Ou seja, dadas p_1 e p_2, duas distribuições de probabilidade, e fixando-se q, a distribuição de referência ou distribuição base. Temos então, para $\lambda \in [0, 1]$,*

$$D(\lambda p_1 + (1-\lambda)p_2|q) \leq D(p_1|q) + (1-\lambda)D(p_2|q). \tag{84}$$

Demonstração.

$$D(\lambda p_1 + (1-\lambda)p_2|q) = \sum_x (\lambda p_1 + (1-\lambda)p_2) \log \frac{\lambda p_1 + (1-\lambda)p_2}{q} \tag{85a}$$

$$\leq \lambda \sum_x p_1 \log \frac{p_1}{q} + (1-\lambda) \sum_x p_2 \log \frac{p_2}{q} \tag{85b}$$

$$= \lambda D(p_1||q) + (1-\lambda)D(p_2||q), \tag{85c}$$

onde em 85a utilizamos que a função $f(p) = p \log p$ é convexa em p (lembrando que q é fixo e assim $f''(p) = {}^1/_p > 0$, para $p > 0$) e utilizando a desigualdade de Jensen (Teorema 6). Em 85b apenas utilizamos a definição de divergência (Definição 5).

□

A entropia relativa, entretanto, não é convexa no segundo argumento. Será entretanto côncava em q para p fixo. A demonstração deixamos a cargo do leitor.

O Teorema 9 pode ser utilizado para demonstrar a concavidade da entropia.

Teorema 10 (A Entropia é Côncava) *$H(p)$ é uma função concava de p.*

Demonstração.

$$H(p) = -\sum_i p_i \log p_i = -\sum_i p_i \log p_i + \log|\mathcal{X}| - \log|\mathcal{X}| \tag{86a}$$

$$= \log|\mathcal{X}| - \sum_i p_i \log p_i - \log|\mathcal{X}| \underbrace{\sum_i p_i}_{=1} \tag{86b}$$

$$= \log|\mathcal{X}| - \sum_i (p_i \log p_i + p_i \log|\mathcal{X}|) \tag{86c}$$

$$= \log|\mathcal{X}| - \sum_i p_i \left(\log p_i - \log {}^1/_{|\mathcal{X}|}\right) \tag{86d}$$

$$= \underbrace{\log|\mathcal{X}|}_{\text{constante}} - \underbrace{D(p||u)}_{\text{convexo}} \tag{86e}$$

□

A partir da relação em 86e, podemos ver a entropia como a similaridade com a distribuição uniforme. Quanto maior a entropia (menor a divergência $D(p||u)$), mais próximo estaremos da distribuição uniforme.

Analisaremos agora a convexidade ou concavidade da informação mútua $I(X;Y)$, quando fixamos uma das distribuições condicionais ou marginais. Essa análise possui implicação direta na determinação da capacidade de canal em teoria da comunicação, onde a maximização da informação mútua, sobre a distribuição de entrada, será um problema bem comportado com um único máximo global. A concavidade da função objetivo implica que o máximo é atingível e pode ser encontrado usando técnicas de otimização convexa.

Teorema 11 (Concavidade/Convexidade da Informação Mútua) *Seja $(X,Y) \sim p(x,y) = p(x)p(y|x)$, a informação mútua $I(X;Y)$ é uma função côncava de $p(x)$ para $p(y|x)$ fixo e uma função convexa de $p(y|x)$ para $p(x)$ fixo.*

Demonstração. Vejamos primeiro o motivo de $I(X;Y)$ ser uma função côncava de $p(x)$ para $p(y|x)$ fixo. Podemos escrever a informação

mútua como uma função de $p(x)$:

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \tag{87a}$$

$$= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{87b}$$

$$= \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}, \tag{87c}$$

onde em 87a utilizamos a definição de informação mútua, em 87b utilizamos a definição de divergência e aplicamos o teorema de Bayes em 87c. Se $p(y|x)$ é constante, então a informação mútua é função de $p(x)$: $I_{p(x)}(X;Y)$. Utilizando a propriedade da convexidade da divergência de Kullback-Leibler,

$$I_{\lambda p_1(x)+(1-\lambda)p_2(x)}(X;Y) \geq \lambda I_{p_1(x)}(X;Y) + (1-\lambda) I_{p_2(x)}(X;Y). \tag{88}$$

Então a informação mútua é uma função concava de $p(x)$ para $p(y|x)$ fixo.

Agora iremos demonstrar que $I(X;Y)$ é uma função convexa de $p(y|x)$ para $p(x)$ fixo. Aplicamos a mesma ideia, porém agora consideraremos $p(x)$ fixo. Escrevemos então a informação mútua como uma função de $p(y|x)$:

$$I_{p(y|x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)}. \tag{89}$$

Utilizando a propriedade da convexidade da divergência de Kullback-Leibler, obtemos

$$I_{\lambda p_1(y|x)+(1-\lambda)p_2(y|x)}(X;Y) \leq \lambda I_{p_1(y|x)}(X;Y) + (1-\lambda) I_{p_2(y|x)}(X;Y). \tag{90}$$

□

Desigualdade de Processamento de Dados

Quando falamos em processamento de dados, podemos pensar no seguinte modelo: $X \to Y \to Z$, onde a variável aleatória X representa os dados originais, a serem transmitidos; Y representa os dados recebidos no outro lado de um canal de comunicação; e Z o resultado de um processamento adicional aplicado a Y. O processamento de dados é modelado como uma cadeia de Markov. Nessa estrutura, Z depende de X apenas através de Y, ou, de outra forma, Z e X são condicionalmente independentes, dado Y.

Definição 8 (Cadeia de Markov)**.** As variáveis aleatórias X, Y e Z formam uma cadeia de Markov nesta ordem (denotado $X \to Y \to Z$) se a distribuição condicional de Z depende apenas de Y e é condicionalmente independente de X. Especificamente, X, Y e Z formam uma cadeia de Markov $X \to Y \to Z$ se a função massa de probabilidade conjunta pode ser escrita como

$$p(x, y, z) = p(x)p(y|x)p(z|y) \tag{91}$$

Temos uma cadeia de Markov $X \to Y \to Z$ se, e somente se, X e Z são condicionalmente independentes dado Y ($X \perp\!\!\!\perp Z|Y$). Isto é,

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = p(x|y)p(z|y) \quad \forall x, y, z \tag{92}$$

Proposição 10 (Cadeia de Markov Inversa) *Se X, Y e Z formam uma cadeia de Markov nesta ordem ($X \to Y \to Z$), então também formam uma cadeia de Markov na ordem inversa ($Z \to Y \to X$).*

Demonstração.

$$p(x, y, z) = p(x)p(y|x)p(z|y) = p(x, y)p(z|y) \tag{93a}$$
$$= \frac{p(x, y)p(z|y)p(y)}{p(y)} = p(x|y)p(y, z) \tag{93b}$$
$$= p(x|y)p(y|z)p(z) \tag{93c}$$

□

A Desigualdade do Processamento de Dados estabelece que a informação mútua entre variáveis aleatórias não aumenta sob o processamento de dados. Em outras palavras, a informação relevante para uma variável não pode ser aumentada ao processar outra variável relacionada em uma cadeia de Markov.

Teorema 12 (Desigualdade do Processamento de Dados) *Se $X \to Y \to Z$ então*

$$I(X; Y) \geq I(X; Z) \tag{94}$$

Demonstração. Utilizando a regra da cadeia da informação mútua (Proposição 6), teremos

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) \tag{95a}$$
$$= I(X; Y) + I(X; Z|Y) \tag{95b}$$

Como $X \perp\!\!\!\perp Z|Y$ (X e Z são condicionalmente independentes, dado Y), temos que $I(X; Z|Y) = 0$. Como $I(X; Y|Z) \geq 0$, usando eqs. (95a) e (95b), teremos

$$I(X; Z) + \underbrace{I(X; Y|Z)}_{\geq 0} = I(X; Y) + \cancelto{0}{I(X; Z|Y)} \tag{96}$$

Podemos conclui então que

$$I(X;Z) \leq I(X;Y). \tag{97}$$

□

Note que teremos igualdade na Equação (94) se, e somente se, $I(X;Y|Z) = 0$ (i.e. $X \to Z \to Y$). De forma similar, também podemos mostrar que $I(Y;Z) \geq I(X;Z)$.

Corolário 12.1 *Se $Z = g(Y)$, então $I(X;Y) \geq I(X;g(Y))$.*

Demonstração. $X \to Y \to g(Y)$ forma uma cadeia de Markov.

□

Corolário 12.2 *Se $X \to Y \to Z$ então $I(X;Y|Z) \leq I(X;Y)$.*

Demonstração.

$$\underbrace{I(X;Z)}_{\geq 0} + I(X;Y|Z) = I(X;Y) + \cancelto{0}{I(X;Z|Y)} \tag{98}$$

então

$$I(X;Y|Z) \leq I(X;Y). \tag{99}$$

□

Note também que

$$\begin{aligned} I(X;Y|Z) &= H(X|Z) - H(X|Y,Z) && \text{(100a)} \\ &\leq H(X) - H(X|Y) && \text{(100b)} \\ &= I(X;Y). && \text{(100c)} \end{aligned}$$

O processamento digital de imagem para remoção de ruído, por exemplo, é uma aplicação prática para a desigualdade de processamento de dados. Considere X como a imagem original, Y como a imagem ruidosa recebida em uma comunicação, e Z a imagem após o processamento para remoção de ruído. Podemos considerar que temos aqui uma cadeia de Markov $X \to Y \to Z$ e assim podemos aplicar a desigualdade de processamento de dados, que nos diz: $I(X;Z) \leq I(X;Y)$. O processo de remoção de ruído não pode aumentar a informação mútua entre a imagem original X e a imagem processada Z além do que já existe entre X e Y, indicando uma perda inevitável de informação devido ao ruído. No entanto, a remoção de ruído é possível porque o algoritmo pode explorar regularidades estatísticas ou estruturais na imagem.

Algumas Relações Notáveis

Proposição 11 (Determinismo e a incerteza nula) $H(X) = 0$ *se, e somente se,* X *for determinístico.*

Demonstração. Se X é determinístico, então existe $x^* \in \mathcal{X}$ tal que $p(x^*) = 1$ e $p(x) = 0 \forall x \neq x^*$. Neste caso, teremos $H(X) = -p(x^*)\log p(x^*) = 0$. Por outro lado, se X não for determinístico, então existe $x^* \in \mathcal{X}$ tal que $0 < p(x^*) < 1$. Desta forma, $H(X) \geq -p(x^*)\log p(x^*) \geq 0$. Podemos concluir assim que $H(X) = 0$ se e somente se X for determinístico.

□

Proposição 12 (Quando Y é uma função de X) $H(Y|X) = 0$ *se, e somente se,* Y *for um função de* X*, ou seja,* $Y = g(X)$.

Demonstração. Observando a Equação (28a), podemos concluir que $H(Y|X) = 0$ se e somente se todos os termos do somatório forem nulos, ou seja, $H(Y|X = x) = 0$ para cada $x \in S_X$. A partir da Proposição 11, sabemos que isto ocorre se, e somente se, Y é determinístico para cada x dado. Em outras palavras, Y é uma função de X.

□

Demonstração. Uma outra forma de demonstração é por contradição. Assuma que existe x, digamos x_0, e dois valores diferentes de y, digamos y_1 e y_2, tal que $p(x_0, y_1) > 0$ e $p(x_0, y_2) > 0$. Então a marginal é $p(x_0) \geq p(x_0, y_1) + p(x_0, y_2) > 0$. Temos também

$$p(y_1|x_0) = \frac{p(x_0, y_1)}{p(x_0)} \text{ e } p(y_2|x_0) = \frac{p(x_0, y_2)}{p(x_0)} \tag{101}$$

então ambos $p(y_1|x_0)$ e $p(y_2|x_0)$ não são iguais a 0 (zero) ou 1 (um).

$$\begin{aligned}
H(Y|X) &= E[H(Y|X)] && \text{(102a)}\\
&= -\sum_x p(x)H(Y|X = x) && \text{(102b)}\\
&= -\sum_x p(x)\sum_y p(y|x)\log p(y|x) && \text{(102c)}\\
&\geq -p(x_0)\sum_y p(y|x_0)\log p(y|x_0) && \text{(102d)}\\
&\geq -\underbrace{p(x_0)}_{>0}[\underbrace{p(y_1|x_0)}_{>0}\underbrace{\log p(y_1|x_0)}_{<0} + \underbrace{p(y_2|x_0)}_{>0}\underbrace{\log p(y_2|x_0)}_{<0}] && \text{(102e)}\\
&> 0 && \text{(102f)}
\end{aligned}$$

Então, a entropia condicional $H(Y|X)$ é nula se e somente se Y for uma função de X. Se Y for uma função de X, teremos $p(y_1|x_0) = 0$ ou 1, ou seja, a probabilidade $p(y_i|x_0)$ será igual a 1 apenas para um y_i e zero para os demais.

□

Teorema 13 (Informação mútua condicional é não nula) *Para variáveis aleatórias X, Y e Z,*

$$I(X;Y|Z) \geq 0, \tag{103}$$

com igualdade se, e somente se, X e Y forem independentes quando condicionados a Z, ou seja, $X \perp\!\!\!\perp Y|Z$.

Demonstração. Observe que

$$I(X;Y|Z) = \sum_{x,y,x} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{104a}$$

$$= \sum_{z} p(z) \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{104b}$$

$$= \sum_{z} p(z) D(p_{XY|z} || p_{X|z} p(Y|z)), \tag{104c}$$

onde $p_{XY|z}$ representa $\{p(x,y|z), (x,y) \in \mathcal{X} \times \mathcal{Y}\}$, $p_{X|z}$ representa $\{p(x|z), x \in \mathcal{X}\}$, e $p_{Y|z}$ representa $\{p(y|z), y \in \mathcal{Y}\}$. Dado z, ambos $p_{XY|z}$ e $p_{X|z}p_{Y|z}$ são distribuições conjuntas em $\mathcal{X} \times \mathcal{Y}$. Teremos então

$$D(p_{XY|z} || p_{X|z} p(Y|z)) \geq 0. \tag{105}$$

Concluímos assim que $I(X;Y|Z) \geq 0$. A partir da Definição 5, observamos que $I(X;Y|Z) = 0$ se, e somente se, para todo $z \in S_z$,

$$p(x,y|z) = p(x|z)p(y|z) \tag{106a}$$

ou

$$p(x,y,z) = p(x,z)p(y|z) \tag{106b}$$

para todo x e y. Então X e Y são independentes quando condicionados a Z.

□

Proposição 13 (Independência e informação mútua nula) $I(X;Y) = 0$ *se, e somente se, $X \perp\!\!\!\perp Y$.*

Demonstração. Este caso equivale ao Teorema 13 com a variável Z degenerada.

□

Teorema 14 (Condicionar não aumenta a entropia)

$$H(Y|X) \leq H(Y) \tag{107}$$

com igualdade se, e somente se, $X \perp\!\!\!\perp Y$.

Demonstração. Basta considerar

$$H(Y|X) = H(Y) - \underbrace{I(X;Y)}_{\geq 0} \leq H(Y). \tag{108}$$

A igualdade ocorrerá se, e somente se, $I(X;Y) = 0$, o que equivale a $X \perp\!\!\!\perp Y$, conforme a Proposição 13.

□

A próxima desigualdade estabelece um teto para a entropia conjunta de um conjunto de variáveis aleatórias, assumindo que elas são independentes.

Teorema 15 (Limite de Independência para Entropia) *A entropia conjunta $H(X_1, X_2, \ldots, X_n)$ de um conjunto de variáveis aleatórias $X_1, X_2, \ldots, X_n$ é sempre menor ou igual à soma das entropias individuais, ou seja,*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i), \tag{109}$$

com igualdade se, e somente se, X_i, $i = 1, 2, \ldots, n$, forem mutuamente independentes.

Demonstração.

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) \tag{110a}$$

$$\leq \sum_{i=1}^{n} H(X_i), \tag{110b}$$

onde em 110a utilizamos a regra da cadeia e em 110b utilizamos que condicionar não aumenta a entropia. A desigualdade é justa se e somente se for justa para cada i em 110b, ou seja,

$$H(X_i|X_1, \ldots, X_{i-1}) = H(X_i), \tag{111}$$

para $i = 1, \ldots, n$. Isto equivale a ter $X_i \perp\!\!\!\perp X_j$, $\forall i \neq j$.

□

Teorema 16

$$I(X; Y, Z) \geq I(X; Y), \tag{112}$$

com igualdade se, e somente se, tivermos uma cadeia de Markov $X \to Y \to Z$.

Demonstração. Pela regra da cadeia da informação mútua obtemos

$$I(X; Y, Z) = I(X; Y) + I(X; Y|Z) \geq I(X; Y), \tag{113}$$

onde, para obter a desigualdade,consideramos o fato de que qualquer informação mútua é não negativa. Teremos $I(X; Y|Z) = 0$ se $X \to Y \to Z$ formar uma cadeia de Markov.

□

Note que o Teorema 16 não contradiz o Corolário 12.2. Quando existir uma cadeia de Markov $X \to Y \to Z$, teremos simplesmente $I(X; Y, Z) = I(X; Y)$, mostrando que Z não adiciona informação sobre X além do que Y já fornece, uma propriedade que reflete a estrutura de dependência da cadeia de Markov.

Veremos agora uma relação importante que conecta a entropia à probabilidade de colisão entre amostras independentes, sendo útil em aplicações como compressão de dados e análise de colisões em funções hash, além de reforçar a propriedade da distribuição uniforme como maximizadora da entropia.

Proposição 14 (Limite inferior da probabilidade de colisão) *Considere duas variáveis aleatórias i.i.d. X e X' com entropia $H(X)$. Teremos o seguinte limite inferior para a probabilidade de colisão $(X = X')$:*

$$Pr(X = X') \geq 2^{-H(X)} \tag{114}$$

com igualdade se e somente se X possuir distribuição uniforme.

Demonstração. Podemos calcular explicitamente o valor probabilidade de coincidência para duas variáveis aleatórias:

$$Pr(X = X') = Pr(X = x_1|X' = x_1)Pr(X' = x_1) + \ldots + Pr(X = x_n|X' = x_n)Pr(X' = x_n) \tag{115a}$$

$$= Pr(X = x_1)Pr(X' = x_1) + \ldots + Pr(X = x_n)Pr(X' = x_n) \tag{115b}$$

$$= p^2(x_1) + \ldots + p^2(x_n) = \sum_x p^2(x). \tag{115c}$$

Em muitas aplicações práticas, a distribuição exata p pode não ser conhecida, mas a entropia pode ser estimada. Torna-se também intuitivo buscar interpretar a probabilidade de coincidência em termos de incerteza. Para tanto, suponha que $X \sim p(x)$. Usando a desigualdade de Jensen temos

$$2^{E[\log p(X)]} \leq E[2^{\log p(X)}], \tag{116}$$

pois 2^x é convexa. Logo,

$$2^{-H(X)} = 2^{\sum_x p(x)\log p(x)} = 2^{E[\log p(X)]} \tag{117a}$$

$$\leq E[2^{\log p(X)}] \tag{117b}$$

$$= \sum_x p(x)2^{\log p(x)} = \sum_x p(x)p(x) \tag{117c}$$

$$= \sum_x p^2(x) = Pr(X = X') \tag{117d}$$

□

Note que, para maximizar a probabilidade $Pr(X = X')$, devemos minimizar a entropia. No limite, quando $H(X) = 0$, teremos $Pr(X = X') \geq 1$, logo será igual a 1 e assim $X = X'$ sem dúvida. Por outro lado, maximizar a entropia garante que este limite inferior será mínimo. Isso apenas nos garante que há margem para minimizar

a probabilidade de colisão até determinado ponto, não restringe superiormente esta problabilidade, o que seria desejavál de se encontrar.

A Equação (115c) fornece a probabilidade de colisão.

Definição 9 (Entropia de Colisão)**.** A chamada entropia de colisão, ou entropia de Rényi de segunda ordem ($n = 2$) é definida com

$$H_2(X) = -\log(Pr(X = X')). \tag{118}$$

$H_2(X)$ mede a incerteza de que duas realizações de dois experimentos aleatórios independentes X e X' tenham o mesmo resultado. Muitas das propriedades básicas da entropia de Shannon não podem ser aplicadas à entropia de Rényi. Para mais informações sobre este tópico, recomendamos a leitura de Delfs e Knebl (2015).

Vamos analisar agora a probabilidade de colisão quando as duas variáveis aleatórias possuem distribuições distintas.

Corolário 16.1 (Limite inferior da probabilidade de colisão com distribuições distintas) *Seja X, X' independentes com $X \sim p(x)$ e $X' \sim q(x)$, $x, x' \in \mathcal{X}$, então*

$$Pr(X = X') \geq 2^{-H(p)-D(p||q)} \tag{119a}$$
$$Pr(X = X') \geq 2^{-H(q)-D(q||p)} \tag{119b}$$

ou seja,

$$Pr(X = X') \geq \max\left(2^{-H(p)-D(p||q)}, 2^{-H(q)-D(q||p)}\right) \tag{120}$$

Demonstração.

$$\begin{aligned}
2^{-H(p)-D(p||q)} &= 2^{\sum_x p(x)\log p(x)+\sum_x p(x)\log\frac{q(x)}{p(x)}} && \text{(121a)}\\
&= 2^{\sum_x p(x)\log q(x)} && \text{(121b)}\\
&= 2^{E_p[\log q(X)]} && \text{(121c)}\\
&\leq \sum_x p(x)2^{\log q(x)} && \text{(121d)}\\
&= \sum_x p(x)q(x) && \text{(121e)}\\
&= Pr(X = X'), && \text{(121f)}
\end{aligned}$$

onde em 121d aplicamos a desigualdade de Jensen.

□

Comunicação em um canal ruidoso

A comunicação em um canal ruidoso é um cenário clássico em teoria da informação, onde uma mensagem X é transmitida através de um canal sujeito a ruído, resultando em uma versão distorcida Y. O receptor, tenta estimar X a partir de Y, gerando uma estimativa $\hat{X} = g(Y)$.

Esse processo forma uma cadeia de Markov: $X \to Y \to \hat{X}$. O processo de estimação está inevitavelmente sujeito a erros Para quantificar esse erro e estabelecer limites sobre a probabilidade de erro na estimação de X, introduziremos a desigualdade de Fano.

Primeiramente, definiremos o erro na estimação como o evento em que $X \neq \hat{X}$, e a probabilidade de erro será denotada por $P_e = p(X \neq \hat{X})$. Em muitas situações, não é possível calcular diretamente P_e; no entanto, veremos que a incerteza sobre X dado Y, representada por $H(X\,Y)$, pode ser usada para estabelecer um limite inferior para essa probabilidade de erro. Para derivar esse limite, será suficiente conhecer as características do canal, ou seja, a distribuição condicional $p(y|x)$, que descreve a relação entre a entrada X e a saída Y.

Teorema 17 (Desigualdade de Fano) *Para qualquer estimador $\hat{X}$ tal que $X \to Y \to \hat{X}$, com $P_e = Pr(X \neq \hat{X})$, temos*

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y). \tag{122}$$

Esta desigualdade pode ser simplificada[10] *na forma*

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)}, \tag{124}$$

onde utilizamos $H(P_e) \leq 1$.

[10] Para o caso de um alfabeto binário ($|\mathcal{X}| = 2$), a desigualdade de Fano na forma da Equação (124) não poderá ser aplicada. Devemos então utilizar a forma mais relaxada:

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)} > \frac{H(X|Y) - 1}{\log|\mathcal{X}|} \tag{123}$$

Demonstração. Vamos primeiramente definir uma função de erro:

$$E = \begin{cases} 1 & , \text{ se } \hat{X} \neq X (\text{erro}) \\ 0 & , \text{ se } \hat{X} = X (\text{sem erro}). \end{cases} \tag{125}$$

Utilizando a regra da cadeia obtemos:

$$H(E, X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} \tag{126a}$$

$$= \underbrace{H(E|\hat{X})}_{\leq H(E) = H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log(|\mathcal{X}|-1)}. \tag{126b}$$

O termo $H(E|X, \hat{X})$ é nulo pois o erro é uma função determinística de X e $\hat{X}$, então, sabendo X e $\hat{X}$, determinamos E. Como condicionar não aumenta a entropia, usamos $H(E|\hat{X}) \leq H(E) = H(P_e)$. E por fim, note que

$$H(X|\hat{X}, E) = p(E = 0)\underbrace{H(X|\hat{X}, E = 0)}_{=0} + p(E = 1)H(X|\hat{X}, E = 1) \tag{127a}$$

$$= (1 - P_e)0 + P_e H(X|\hat{X}, E = 1) \leq P_e \log(|\mathcal{X}| - 1), \tag{127b}$$

onde usamos que, se não há erro ($E = 0$) e conhecemos $\hat{X}$, então determinamos X. Não existe entropia residual em X quando é dado $\hat{X}$ e $E = 0$. Logo $H(X|\hat{X}, E = 0) = 0$. Usamos ainda que se conhecemos $\hat{X}$ e existe um erro ($E = 1$), então sabemos que X é diferente de $\hat{X}$, logo isto nos deixa com $(|\mathcal{X}| - 1)$ alternativas. Assim, teremos $H(X|\hat{X}, E = 1) \leq \log(|\mathcal{X}|-1)$. Aplicando $H(X|\hat{X}, E) \leq P_e \log(|\mathcal{X}|-1)$ na Equação (126), obtemos agora

$$H(X|\hat{X}) = H(E|\hat{X}) + H(X|E, \hat{X}) \quad \leq H(P_e) + P_e \log(|\mathcal{X}| - 1). \tag{128a}$$

Como $X \to Y \to \hat{X}$ é uma cadeia de Markov, podemos utilizar a desigualdade de processamento de dados:

$$I(X;Y) \geq I(X;\hat{X}) \tag{129a}$$
$$H(X) - H(X|Y) \geq H(X) - H(X|\hat{X}) \tag{129b}$$
$$H(X|\hat{X}) \geq H(X|Y) \tag{129c}$$

Então, utilizando as Equações 128 e 129, obtemos

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y). \tag{130}$$

□

Exemplo 4. Considere uma variável aleatória discreta $X \in \mathcal{X} = \{1, 2, \ldots, 5\}$ com função massa de probabilidade $p(x) = (0.35, 0.35, 0.1, 0.1, 0.1)$. Seja $Y \in \mathcal{Y} = \{1, 2\}$, de forma que, se $x \leq 2$ teremos $y = x$ com probabilidade $6/7$ e, se $x > 2$, teremos $y = 1$ ou 2 com igual probabilidade. A melhor estratégia é utilizar o estimador $\hat{x} = y$. Calcule a probabilidade de erro e o limite dado pela desigualdade de Fano.

A distribuição condicional $p(y|x)$ é

X \ Y	1	2
1	6/7	1/7
2	1/7	6/7
3	1/2	1/2
4	1/2	1/2
5	1/2	1/2

A efetiva probabilidade de erro é dada por

$$P_e = 1 - P_a \tag{131a}$$
$$= 1 - \sum_{i=1}^{5} P(x_i = y_i) \tag{131b}$$
$$= 1 - (p(y = 1|x = 1)p(x = 1) + p(y = 2|x = 2)p(x = 2) + 0 + 0 + 0) \tag{131c}$$
$$= 1 - \left(\frac{6}{7}0.35 + \frac{6}{7}0.35\right) = 0.4 = \frac{2}{5} \tag{131d}$$

onde denotamos por P_a a probabilidade de acerto.

A desigualdade de Fano fornece um limite inferior pra a probabilidade de erro (predição incorreta do valor de X baseado em Y). Este limite inferior é determinado pela incerteza remanescente $H(X|Y)$ sobre X quando Y é conhecido.

Pelo teorema da desigualdade de Fano temos que

$$P_e \geq \frac{H(X|Y)-1}{\log\left(|\mathcal{X}|-1\right)} \tag{132}$$

$$\begin{aligned} H(X|Y) &= -\sum_{x,y} p(x,y)\log p(x|y) && \text{(133a)}\\ &= -\sum_{x,y} p(y|x)p(x)\log p(x|y) && \text{(133b)} \end{aligned}$$

onde $p(y|x)$ e $p(x)$ são dados do problema e ainda será necessário calcular $p(x|y)$ para encontrar $H(X|Y)$.

$$\begin{aligned} P(X|Y=1) &= \frac{P(X,Y=1)}{P(Y=1)} = \frac{P(Y=1|X)P(X)}{P(Y=1)} && \text{(134a)}\\ &= \frac{(\frac{6}{7},\frac{1}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)}{\sum\left((\frac{6}{7},\frac{1}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)\right)} && \text{(134b)}\\ &= \frac{(0.3,0.05,0.05,0.05,0.05)}{1/2} && \text{(134c)}\\ &= (0.6,0.1,0.1,0.1,0.1) && \text{(134d)} \end{aligned}$$

$$\begin{aligned} P(X|Y=2) &= \frac{(\frac{1}{7},\frac{6}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)}{\sum\left((\frac{1}{7},\frac{6}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)\right)} && \text{(135a)}\\ &= (0.1,0.6,0.1,0.1,0.1) && \text{(135b)} \end{aligned}$$

Desta forma Teremos

$$\begin{aligned} H(X|Y) &= H(X|Y=1)P(Y=1) + H(X|Y=2)P(Y=2) && \text{(136a)}\\ &= -\frac{1}{2}\left(0.6\log 0.6 + 4\times 0.1\log 0.1\right) - \frac{1}{2}\left(4\times 0.1\log 0.1 + 0.6\log 0.6\right) && \text{(136b)}\\ &= 1.771 \text{ bits.} && \text{(136c)} \end{aligned}$$

Utilizando a desigualdade de Fano

$$P_e \geq \frac{H(X|Y)-1}{\log\left(|\mathcal{X}|-1\right)} = \frac{1.771-1}{\log(5-1)} = 0.3855. \tag{137}$$

De fato, temos $P_e = 0.4 \geq 0.3855$.

Medida de Informação

Exploramos conceitos como entropia, informação mútua e divergência de Kullback-Leibler sob a perspectiva de variáveis aleatórias e suas distribuições, mas essas grandezas podem ser entendidas de forma mais profunda através de um paralelo com a teoria dos conjuntos e a teoria da medida. Desta forma, buscaremos uma interpretação geométrica e estrutural para relações como a regra da cadeia e a desigualdade do processamento de dados, permitindo uma compreensão mais ampla de como a informação é estruturada e quantificada em espaços probabilísticos.

Dado um conjunto de variáveis aleatórias: $X_1, X_2, \ldots, X_n$, a cada uma delas associamos um conjunto $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

Definição 10 (Campo)**.** Um campo $\mathcal{F}_n$ gerado pelos conjuntos $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ é a coleção de conjuntos que podem ser obtidos através de qualquer sequência de operações usuais de conjuntos (união, interseção, complemento, e diferença) sobre os conjuntos $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

Definição 11 (Átomo)**.** Os átomos de $\mathcal{F}_n$ são os conjuntos da forma $\cap_{i=1}^{n} Y_i$, onde Y_i é $\tilde{X}_i$ ou $\tilde{X}_i^c$, o complemento de $\tilde{X}_i$.

Exemplo 5. Vamos considerar o seguinte exemplo com $n = 2$. Neste caso, teremos os conjuntos $\tilde{X}_1, \tilde{X}_2$ e seus complementos, respectivamente, $\tilde{X}_1^c, \tilde{X}_2^c$. Existirão 4 átomos: $\tilde{X}_1 \cap \tilde{X}_2$, $\tilde{X}_1 \cap \tilde{X}_2^c$, $\tilde{X}_1^c \cap \tilde{X}_2$, e $\tilde{X}_1^c \cap \tilde{X}_2^c$, que estão representados na Figura 10.

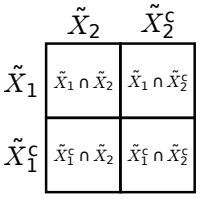

Figura 10: Exemplo de átomos para $n = 2$.

Exemplo 6. Considerando agora o caso com $n = 3$. Teremos os conjuntos $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3$ e seus complementos, respectivamente, $\tilde{X}_1^c, \tilde{X}_2^c, \tilde{X}_3^c$. Existirão 8 átomos: $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3$, $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c$, $\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3$, $\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3^c$, $\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3$, $\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3^c$, $\tilde{X}_1^c \cap \tilde{X}_2^c \cap \tilde{X}_3$, e $\tilde{X}_1^c \cap \tilde{X}_2^c \cap \tilde{X}_3^c$, como representados na Figura 11.

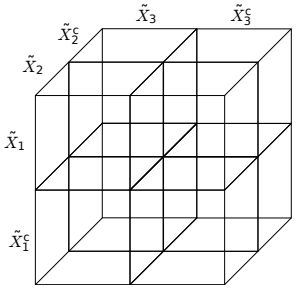

Figura 11: Exemplo de átomos para $n = 3$.

Consideremos n variáveis aleatórias discretas $X_1, X_2, \ldots, X_n$, onde cada X_i tem um suporte finito S_{X_i}, com $|S_{X_i}| = k_i$, e os suportes S_{X_i} podem ser distintos. Por exemplo, $S_{X_1} = \{0, 1\}$, $S_{X_2} = \{a, b, c\}$, e assim por diante. O espaço amostral é o produto cartesiano $\Omega = S_{X_1} \times S_{X_2} \times \ldots \times S_{X_n}$, com $|\Omega| = k_1 \cdot k_2 \cdot \ldots \cdot k_n$, e o campo $F_n = 2^{\Omega}$ contém todos os subconjuntos de Ω. Nesse contexto, os átomos são os eventos $\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\}$, e grandezas como a entropia $H(X_1, \ldots, X_n)$ podem ser interpretadas como medidas sobre F_n. Para ilustrar propriedades específicas, consideremos o caso particular em que $S_{X_i} = \{0, 1\}, \forall i$, ou seja, $\Omega = \{0, 1\}^n$, com $|\Omega| = 2^n$ e $F_n = 2^{\Omega}$. Nesse cenário, teremos as seguintes propriedades:

1. há exatamente 2^n átomos;

2. $|F_n| = 2^{\Omega} = 2^{2^n}$;

3. Todos os átomos em F_n são disjuntos;

4. Todo conjunto $A \in F_n$ pode ser expresso de forma única como uma união de um subconjunto dos átomos.

Em análise matemática, uma medida em um conjunto S é uma forma sistemática de atribuir números a todo subconjunto de S, sendo intuitivamente interpretada como o seu tamanho. Medida com sinal é uma generalização do conceito de medida permitindo que esta assuma valores negativos.

Definição 12 (Medida com sinal)**.** Uma função real μ definida em $\mathcal{F}_n$ é chamada medida com sinal se for aditiva no conjunto, i.e., para A e B disjuntos em $\mathcal{F}_n$,

$$\mu(A \cup B) = \mu(A) + \mu(B). \tag{138}$$

Para uma medida com sinal μ teremos $\mu(\emptyset) = 0$, já que $\mu(A) = \mu(A \cup \emptyset) = \mu(A) + \mu(\emptyset)$.

Uma medida com sinal μ em $\mathcal{F}_n$ é completamente especificada por seus valores nos átomos de $\mathcal{F}_n$. Os valores de μ em outros conjuntos de $\mathcal{F}_n$ podem ser obtidos pela aditividade de conjuntos, já que qualquer $\tilde{X} \in \mathcal{F}_n$ pode ser representado como $\tilde{X} = \cup_{i=1} Y_i$, onde Y_i são átomos escolhidos apropriadamente.

Exemplo 7 ($n = 2$)**.** Uma medida com sinal μ em $\mathcal{F}_2$ é completamente especificada pelos valores $\mu(\tilde{X}_1 \cap \tilde{X}_2)$, $\mu(\tilde{X}_1 \cap \tilde{X}_2^c)$, $\mu(\tilde{X}_1^c \cap \tilde{X}_2)$, e $\mu(\tilde{X}_1^c \cap \tilde{X}_2^c)$.

O valor de μ em $\tilde{X}_1$ pode ser obtido da seguinte forma

$$\begin{aligned} \mu(\tilde{X}_1) &= \mu((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) && \text{(139a)} \\ &= \mu(\tilde{X}_1 \cap \tilde{X}_2) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c). && \text{(139b)} \end{aligned}$$

O valor de μ em $\tilde{X}_1 \setminus \tilde{X}_2$ é dado por

$$\mu(\tilde{X}_1 \setminus \tilde{X}_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c). \tag{140}$$

O valor de μ em $\tilde{X}_1 \cup \tilde{X}_2$ pode ser obtido através de

$$\begin{aligned} \mu(\tilde{X}_1 \cup \tilde{X}_2) = &\quad \mu((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c) \cup (\tilde{X}_1^c \cap \tilde{X}_2)) && \text{(141a)} \\ = &\quad \mu(\tilde{X}_1 \cap \tilde{X}_2) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1^c \cap \tilde{X}_2) && \text{(141b)} \end{aligned}$$

Os conjuntos $\tilde{X}_1$ e $\tilde{X}_2$ estão associados às variáveis aleatória X_1 e X_2. O campo $\mathcal{F}_2$ é gerado por $\tilde{X}_1$ e $\tilde{X}_2$, através dos átomos $(\tilde{X}_1 \cap \tilde{X}_2)$, $(\tilde{X}_1 \cap \tilde{X}_2^c)$, $(\tilde{X}_1^c \cap \tilde{X}_2)$, e $(\tilde{X}_1^c \cap \tilde{X}_2^c)$. O diagrama de informação é apresentado na Figura 12.

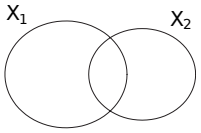

Figura 12: Diagrama de informação para X_1 e X_2.

O conjunto universo será considerada como sendo $\Omega = \tilde{X}_1 \cup \tilde{X}_2$. Desta forma, o átomo $\tilde{X}_1^c \cap \tilde{X}_2^c$ se degenera ao conjunto vazio,

$$\tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \Omega^c = \emptyset. \tag{142}$$

Para as v.a.s X_1 e X_2, as medidas de informação de Shannon são

$$H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2), I(X_1; X_2). \quad (143)$$

Utilizando a notação $A \cap B^c \equiv A \setminus B$, definimos uma medida com sinal μ^*

$$\mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2), \quad (144a)$$
$$\mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1), \quad (144b)$$
$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2). \quad (144c)$$

Estes são os valores de μ^* nos átomos não vazios de $\mathcal{F}_2$. Os valores de μ^* nos demais conjuntos de $\mathcal{F}_2$ podem ser obtidos por adição de conjuntos. Em particular, temos as relações

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2), \quad (145a)$$
$$\mu^*(\tilde{X}_1) = H(X_1), \quad (145b)$$
$$\mu^*(\tilde{X}_2) = H(X_2). \quad (145c)$$

Por exemplo, a Equação 145a pode ser verificada

$$\begin{aligned}\mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= \mu^*((\tilde{X}_1 \setminus \tilde{X}_2) \cup (\tilde{X}_2 \setminus \tilde{X}_1) \cup (\tilde{X}_1 \cap \tilde{X}_2)) \\ &= \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) + \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) + \mu^*(\tilde{X}_1 \cap \tilde{X}_2) \\ &= H(X_1|X_2) + H(X_2|X_1) + I(X_1; X_2) \\ &= H(X_1, X_2). \end{aligned} \quad (146a)$$

A Equação 145b também pode ser facilmente verificada

$$\begin{aligned}\mu^*(\tilde{X}_1) &= \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \\ &= \mu^*(\tilde{X}_1 \cap \tilde{X}_2) + \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) \\ &= I(X_1; X_2) + H(X_1|X_2) = H(X_1). \end{aligned} \quad (147a)$$

É possível então verificar a seguinte correspondência com as medidas de informação de Shannon

$$H/I \leftrightarrow \mu^* \quad (148a)$$
$$, \leftrightarrow \cup \quad (148b)$$
$$; \leftrightarrow \cap \quad (148c)$$
$$| \leftrightarrow \setminus \quad (148d)$$

Com a notação de medida, não existe distinção entre H e I, podemos escrever $H(X;Y) = I(X;Y)$, utilizando a notação do ponto-e-vírgula.

Propriedade da Equipartição Assintótica

A Propriedade da Equipartição Assintótica (AEP) visa analisar o comportamento de sequências no limite, quando estas sequências tornam-se muito grandes. A AEP mostra que para uma sequência longa de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), a probabilidade de observar uma sequência típica é aproximadamente $2^{-nH(X)}$, onde $H(X)$ é a entropia da fonte e n o comprimento da sequência.

Primeiramente, vamos rever o conceito de estatística de uma amostra.

Definição 13 (Estatístima amostral)**.** Seja $x_1, x_2, \ldots, x_n$ uma sequência de comprimento n de valores observados de uma amostra de tamanho n, obtidos a partir da realização de variáveis aleatórias $X_1, X_2, \ldots, X_n$, uma estatística de uma amostra é qualquer função calculada a partir de uma amostra de dados, $T(x_1, x_2, \ldots, x_n)$.

Exemplos comuns incluem a média amostral

$$\bar{x} = T(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{149}$$

a variância amostral

$$s^2 = T(x_1, \ldots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{150}$$

a mediana amostral, a primeira amostra x_1, a última amostra x_n, ou até estatísticas como o máximo ou mínimo da amostra, que capturam aspectos específicos dos dados observados. Quando avaliamos T nas variáveis aleatórias $X_1, X_2, \ldots, X_n$, o resultado $T(X_1, X_2, \ldots, X_n)$ é uma variável aleatória, com sua própria distribuição.

Exemplo 8 (Ensaio de Bernoulli)**.** Considere o experimento de Bernoulli com $X_1, \ldots, X_n$ i.i.d., com $X_i \in \{0, 1\}$ e parâmetro $\theta = Pr(X_i == 1)$.

Uma dada sequência qualquer $x_1, \ldots, x_n$ terá então probabilidade dada por

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \tag{151}$$

Considere agora a seguinte estatística

$$T(x_1, \ldots, x_n) = \sum_{i=1}^{x} x_i, \tag{152}$$

o somatório da amostra, no caso, a quantidade de valoes iguais a 1 que apareceu em uma realização. Uma vez que sabemos tal estatísticas, a probabilidade de uma sequência pode ser expressa sem fazer referência à θ (parâmetro que caracteriza a distribuição).

$$p(x_1, \ldots, x_n | T(x_1, \ldots, x_n), \theta) = p(x_1, \ldots, x_n | T(x_1, \ldots, x_n)) \tag{153a}$$

$$= \begin{cases} \frac{1}{\binom{n}{k}} & , \sum_i x_i = k \\ 0 & , \text{caso contrário.} \end{cases} \tag{153b}$$

Em outras palavras, $X_{1:N} \perp\!\!\!\perp \theta | T(X_{1:N})$. Isto implica na cadeia de Markov $\theta \to T(X_{1:N}) \to X_{1:N}$. Aplicando a desigualdade de processamento de dados, obtemos

$$I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N}). \tag{154}$$

Por outro lado, sabemos que $T(X_{1:N})$ é uma função de $X_{1:N}$. Desta forma, também temos a seguinte cadeia de Markov: $\theta \to X_{1:N} \to T(X_{1:N})$. Novamente, aplicando a desigualdade de processamento de dados, obtemos

$$I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N})). \tag{155}$$

Então, para que 154 e 155 sejam satisfeitos, devemos ter

$$I(\theta; X_{1:N}) = I(\theta; T(X_{1:N})), \tag{156}$$

e nenhuma informação é perdida sobre θ indo de $X_{1:N}$ para $T(X_{1:N})$.

Definição 14 (Estatística Suficiente)**.** Uma função $T(\cdot)$ é dita ser uma estatística suficiente em relação à família $\{f_\theta(x)\}$ se X é independente de θ dado $T(X)$ para qualquer distribuição em θ (i.e. $\theta \to T(X) \to X$ forma uma cadeia de Markov). Então

$$I(\theta; X) = I(\theta; T(X)), \ \forall \theta \tag{157}$$

Uma estatística suficiente preserva a informação mútua e reciprocamente

$$X \perp\!\!\!\perp \theta | T(X). \tag{158}$$

Podemos verificar que, no Exemplo 8, a estatísticas utilizada é uma estatísticas suficiente (veja a Equação (153)).

Um critério prático para identificar estatísticas suficientes é dado pelo Teorema da Fatoração de Fisher-Neyman.

Teorema 18 *Uma estatísticas $T(X_1, \ldots, X_N)$ é suficiente para θ se, e somente se, a função de verossimilhança da amostra pode ser escrita como:*

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = g(T(x_1, \ldots, x_n), \theta) \cdot h(x_1, \ldots, x_n), \tag{159}$$

onde g é uma função que depende de θ e da estatística T, e h é uma função que não depende de θ.

A demonstração e mais informações sobre o tema podem ser vistos em Berger e Casella (2002).

O histograma empírico da amostra é uma estatística que descreve a distribuição de frequências relativas dos valores observados em uma amostra.

Definição 15 (Histograma Empírico)**.** Para uma amostra $x_1, x_2, \ldots, x_n$, obtida a partir de variáveis aleatórias $X_1, X_2, \ldots, X_n$, com suporte finito $S_X = \{a_1, a_2, \ldots, a_D\}$, o histograma empírico pode ser definido como a função que associa cada símbolo $a \in S_X$ à sua frequência relativa:

$$\hat{p}_n(a) = \frac{1}{n} \sum_{i=1}^{n} I\{x_i = a\}, \tag{160}$$

onde $I\{x_i = a\}$ é a função indicadora

$$I\{x_i = a\} = \begin{cases} 1, & x = a, \\ 0, & x \neq a. \end{cases} \tag{161}$$

Desta forma, $\hat{p}_n(a)$ representa a proporção de vezes que o valor a aparece na amostra.

Equivalentemente, o histograma empírico pode ser representado como a enúpla que contém as frequências relativas de todos os símbolos de S_X.

$$P_{x_{1:n}} \triangleq \left(\frac{N(a_1|x_{1:n})}{n}, \frac{N(a_2|x_{1:n})}{n}, \ldots, \frac{N(a_D|x_{1:n})}{n} \right), \tag{162}$$

onde $N(a_i|x_{1:n})$ é a contagem do número de ocorrências do símbolo a_i na amostra $x_{1:n}$. O histograma é uma estatística, já que é uma função da amostra. Podemos mostrar ainda que o histograma empírico é uma estatísticas suficiente:

$$p(x_{1:n}|P_{x_{1:n}}, \theta) = \begin{cases} \frac{1}{\binom{n}{n_1, n_2, \ldots, n_D}} & , n_i = nP_{x_{1:n}}(a_i), \forall i \\ 0 & , \text{caso contrário} \end{cases} \tag{163a}$$

$$= p(x_{1:n}|P_{x_{1:n}}) \tag{163b}$$

onde temos o coeficiente multinomial[11] $\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \ldots k_m!}$. Podemos observar que $p(x_{1:n}|P_{x_{1:n}}, \theta) = p(x_{1:n}|P_{x_{1:n}})$, ou seja, é independente de θ. Então $X_{1:n} \perp\!\!\!\perp \theta | P_{x_{1:n}}$, então $P_{x_{1:n}}$ é uma estatística suficiente.

[11] Teorema Multinomial

$$(x_1 + x_2 + \ldots + x_m)^n = \sum_{k_1 + k_2 + \ldots + k_m = n} \binom{n}{k_1, k_2, \ldots, k_m} \tag{164}$$

Codificação

O codificador é responsável por associar cada sequência $x_1, x_2, \ldots, x_n$ produzida pela fonte, ou seja, realizações de variáveis aleatórias $X_1, X_2, \ldots, X_n$, a uma sequência de bits de comprimento variável ou fixo, de modo que a sequência original possa ser recuperada perfeitamente por um decodificador. Para simplificar, vamos supor que as mensagens codificadas possuem, todas elas, um mesmo comprimento m. O alfabeto da fonte é $\mathcal{X} = \{a_1, a_2, \ldots, a_K\}$, possuindo assim cardinalidade $K = |\mathcal{X}|$. O alfabeto de saída do codificador é $\mathcal{Y} = \{0, 1\}$, ou seja, codificação binária ($|\mathcal{Y}| = 2$). A Figura 13 representa este esquema de codificação.

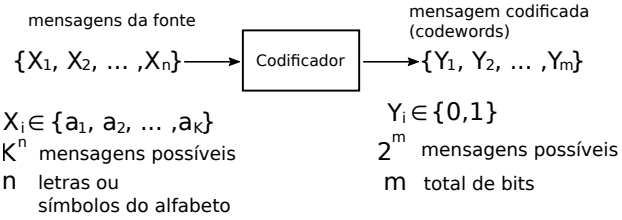

Figura 13: Esquemático de um codificador.

Para que seja possível termos uma palavra de código para cada mensagem possível, devemos satisfazer a seguinte condição:

$$2^m \geq K^n, \tag{165}$$

ou seja,

$$m \geq (\log K)n \tag{166}$$

A taxa de codificação mede a eficiência de um esquema de codificação ao representar uma sequência de símbolos em forma binária.

Definição 16 (Taxa de Codificação)**.** Para uma sequência de entrada de comprimento n, $X_1, X_2, \ldots, X_n$, codificada em uma mensagem binária de comprimento m, $Y_1, Y_2, \ldots, Y_m$, a taxa de codificação é definida como

$$R = \frac{m}{n}. \tag{167}$$

R representa então o número de bits por símbolo utilizados na tarefa de codificação.

Considerando que a fonte produz uma sequência de símbolos i.i.d., a probabilidade de uma sequência qualquer de comprimento n pode ser expressa por

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \tag{168}$$

A informação sobre um evento é dada por $-\log p(x) = I(x)$, então a informação associada à mensagem $x_1, x_2, \ldots, x_n$ é

$$I(x_1, x_2, \ldots, x_n) = -\log p(x_1, x_2, \ldots, x_n) = -\log \prod_{i=1}^{n} p(x_i) \tag{169a}$$

$$= \sum_{i=1}^{n} -\log p(x_i) = \sum_{i=1}^{n} I(x_i), \tag{169b}$$

onde notamos que eventos independentes são aditivos em relação a esta função de informação. Observe que $EI(X) = H(X)$. A lei fraca dos grandes números[12] diz que $\frac{1}{n} S_n \xrightarrow{p} \mu$, onde S_n é a soma de v.a.s i.i.d. com média $\mu = \mathrm{E}X_i$. Temos que $I(X_i)$ também é uma v.a. com média $H(X)$. Obteremos assim

$$\frac{1}{n} \sum_{i=1}^{n} I(X_i) \xrightarrow[n\to\infty]{p} H(X). \tag{171}$$

Quando n é grande suficiente, podemos escrever

$$\frac{1}{n} \sum_{i=1}^{n} I(x_i) \approx H(X), \forall i, x_i \sim p(x) \tag{172a}$$

$$-\frac{1}{n} \sum_{i=1}^{n} \log p(x_i) \approx H(X) \tag{172b}$$

$$-\log \prod_{i=1}^{n} p(x_i) \approx nH(X) \tag{172c}$$

$$-\log p(x_1, x_2, \ldots, x_n) \approx nH(X) \tag{172d}$$

$$p(x_1, x_2, \ldots, x_n) \approx 2^{-nH(X)}. \tag{172e}$$

Esta probabilidade não depende da sequência em si. Depende apenas do comprimento n e da entropia da fonte. Quando n fica grande, podemos dizer que todas as sequências terão a mesma probabilidade: 2^{-nH}. Estas sequências que possuem esta probabilidade (praticamente todas as sequências) são chamadas de *sequências típicas*, e são representadas pela conjunto $A_\epsilon^{(n)}$.

Se todas as sequência de comprimento n possuem aproximadamente a mesma probabilidade $2^{-nH(X)}$, então existe no máximo 2^{nH} sequências de comprimento n. Pode ser que $2^{nH} \ll K^n$, ou seja, o conjunto das sequências típicas é muito menor do que o conjunto de todas as sequências possíveis de comprimento n. Isto implica que, efetivamente, poderemos nos preocupar apenas com a codificação do conjunto menor, o conjunto das sequências que efetivamente ocorrem, as sequências típicas. Desta forma, para representar (ou codificar) as sequencias típicas, precisaremos de $nH(X)$ bits. Teremos então

$$m = nH(X) \tag{173}$$

no modelo do codificador. Então a taxa será $H(X)$.

[12] Lei dos Grandes Números: Se um evento de probabilidade p é observado repetidamente em ocasiões independentes, a proporção da frequência observada deste evento em relação ao total número de repetições converge em direção a p à medida que o número de repetições se torna arbitrariamente grande.

Sejam $X_1, X_2, \ldots, X_n$ v.a.s i.i.d. com $\mathrm{E}X_i = \mu$ e $\mathrm{Var}X_i = \sigma^2 < \infty$, para $i = 1, \ldots, n$. Seja a média definida por $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, então, para $\varepsilon > 0$, a *Lei Fraca dos Grandes Números* diz que $\overline{X_n}$ converge em probabilidade para μ, ou seja,

$$\lim_{n\to\infty} P\left(|\overline{X_n} - \mu| < \varepsilon\right) = 1. \tag{170}$$

Teorema 19 (Propriedade da Equipartição Assintótica) *Se $X_1, X_2, \ldots, X_n$ são i.i.d. e $X_i \sim p(x)$ para todo i, então*

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X) \tag{174}$$

Demonstração.

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) = -\frac{1}{n}\log \prod_{i=1}^{n} p(X_i) \tag{175a}$$

$$= -\frac{1}{n}\sum_i \log p(X_i) \xrightarrow{p} \mathrm{E}\log p(X) \tag{175b}$$

$$= H(X) \tag{175c}$$

onde utilizamos a lei fraca dos números grandes em 175b.

□

Definição 17 (Conjunto Típico)**.** Um conjunto típico $A_\epsilon^{(n)}$ em relação a $p(x)$ é o conjunto de sequências $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ com propriedade

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{176}$$

De forma equivalente, podemos escrever

$$A_\epsilon^{(n)} = \left\{(x_1, x_2, \ldots, x_n) : |-\frac{1}{n}\log p(x_1, \ldots, x_n) - H| < \epsilon\right\} \tag{177}$$

Teorema 20 (Propriedades do Conjunto Típico $A_\epsilon^{(n)}$)

1. *Se $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, então*

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \tag{178}$$

2. *$p(A_\epsilon^{(n)}) = p\left(\left\{x : x \in A_\epsilon^{(n)}\right\}\right) > 1 - \epsilon$ para n grande suficiente, para todo $\epsilon > 0$.*

3. *Limite superior: $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, onde $|A_\epsilon^{(n)}|$ é o número de elementos no conjunto $A_\epsilon^{(n)}$.*

4. *Limite inferior: $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ para n grande suficiente.*

Demonstração.

1. O primeiro apenas é uma reformulação da definição de AEP.

2. Vamos utilizar a definição expandida de convergência em probabilidade, dada na equação 174.

$$p(A_\epsilon^{(n)}) = p\left(|-\frac{1}{n}\sum_i \log p(x_i) - H| < \epsilon\right) > 1 - \delta \tag{179}$$

para n grande suficiente. Podemos escolher qualquer δ, escolhemos então $\delta = \epsilon$, resultando em

$$p(A_\epsilon^{(n)}) > 1 - \epsilon, \quad \text{para } n \text{ grande suficiente } \forall \epsilon \tag{180}$$

3. Limite superior de $A_\epsilon^{(n)}$

$$1 = \sum_x p(x) \geq \sum_{x \in A_\epsilon^{(n)}} p(x) \geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \tag{181a}$$

$$= |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)} \tag{181b}$$

Resultando em $|A_\epsilon^{(n)}| \leq 2^{n(H+\epsilon)}$.

4. Limite inferior do tamanho de $A_\epsilon^{(n)}$. Para n grande suficiente

$$1 - \epsilon < p(A_\epsilon^{(n)}) \leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \tag{182a}$$

$$= 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}| \tag{182b}$$

resultando em $|A_\epsilon^{(n)}| \geq (1 - \epsilon) 2^{n(H(X)-\epsilon)}$.

□

A AEP e o tamanho do conjunto típico têm implicações diretas na codificação de sequências, como no exemplo de imagens digitais. Considere uma imagem full HD de 1080×720 pixels, com 16 milhões de cores (24 bits por pixel, ou seja, 2^{24} cores possíveis). O número total de pixels é $1080 \times 720 = 777.600$, e cada pixel é representado por 24 bits, totalizando $K = 1080 \times 720 \times 24 = 18.662.400 \approx 10^7$ bits por imagem. O número total de imagens possíveis é $2^K = 2^{18.662.400} \approx 10^{5.617 \times 10^6}$, um valor extremamente grande, muito maior que o número de átomos no universo observável ($\approx 10^{81}$). Pela AEP, se os pixels fossem i.i.d. com entropia $H(X)$ (em bits por pixel), o número de imagens típicas seria aproximadamente $2^{nH(X)}$, onde $n = 777.600$. Mesmo que $H(X)$ fosse pequeno (por exemplo, 1 bit por pixel devido a redundâncias), o número de imagens típicas seria $2^{777.600} \approx 10^{234.000}$, ainda muito maior que 10^{81}. Isso ilustra que, embora o número de imagens típicas seja uma fração minúscula do total, ele ainda é astronomicamente grande, destacando a necessidade de codificação eficiente para lidar com sequências típicas, que dominam a probabilidade.

Considere agora as chaves privadas de Bitcoin. Uma chave privada de Bitcoin é um número de 256 bits, gerado aleatoriamente, o que significa que o número total de chaves possíveis é $2^{256} \approx 1,1579 \times 10^{77}$. Esse valor é menor que o número de átomos no universo observável, mas ainda é astronomicamente grande. Pela AEP, se os bits da chave fossem gerados de forma i.i.d. com entropia $H(X)$ (em bits por bit),

o número de chaves típicas seria aproximadamente $2^{nH(X)}$, onde $n = 256$. No caso ideal, em que os bits são uniformes ($H(X) = 1$ bit por bit), o número de chaves típicas é $2^{256 \cdot 1} = 2^{256}$, ou seja, todas as chaves são típicas. No entanto, se houvesse algum viés na geração das chaves, por exemplo, $H(X) = 0,9$ bits por bit, o número de chaves típicas seria $2^{256 \cdot 0,9} = 2^{230,4} \approx 10^{69,3}$, ainda um número enorme, mas uma fração do total. Isso ilustra a importância de maximizar a entropia na geração de chaves para garantir que o espaço de chaves típicas seja o maior possível, minimizando a probabilidade de colisões.

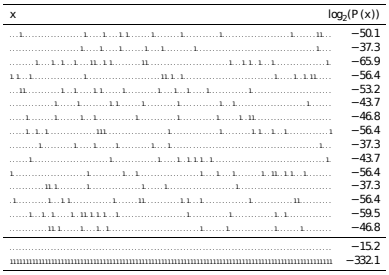

Figura 14: Sequências regadas por um ensaio de Bernoulli com $n = 100$ e $P(X = 1) = p = 0.1$. As 15 sequências superiores representam amostras típicas. As duas últimas sequências representam a sequência mais provável e a menos provável (MacKay 2003).

Considere agora um ensaio de Bernoulli com $n = 100$ lançamentos, onde cada lançamento X_i é uma variável aleatória com $P(X_i = 1) = p = 0,1$ e $P(X_i = 0) = 1 - p = 0,9$. Nesse cenário, uma sequência $(x_1, x_2, \ldots, x_{100}) \in \{0,1\}^{100}$ tem probabilidade $P(x_1, \ldots, x_{100}) = p^k(1-p)^{n-k}$, onde $k = \sum_{i=1}^{100} x_i$ é o número de 1's na sequência. A Figura 14 compara três tipos de sequências: a mais provável, a menos provável e as sequências típicas. A sequência mais provável é a sequência com 100 ocorrências de 0 (penúltima sequência apresentada na Figura 14), já a sequência menos provável é aquela com 100 ocorrências de 1 (última sequência apresentada na Figura 14). As sequências típicas são aquelas com probabilidade aproximadamente $2^{-nH(X)}$. Para o exemplo em questão, temos $H = 0,4690$ e assim $\log(2^{-100H(X)}) = -46.9$. As 15 sequências apresentadas no parte superior da Figura 14 representam aquelas com $\log p(x_1, \ldots, x_n)$ próximo de -46.9, ou seja, sequências que podemos considerar típicas (dado um determinado critério ϵ).

A Figura 15 ilustra o que ocorre quando n cresce, fazendo assim uma comparação para os casos $n = 100$ e $n = 1000$. Primeiramente, temos o número de sequências de comprimento n com r ocorrências de 1's. Este número é dado pelo coeficiente binomial. Nos extremos, existe apenas uma sequência com 100 zeros e uma sequência com 100 1's. No meio, teremos o pico, sequências com 50 ocorrências de cada símbolo. Logo abaixo apresenta-se a probabilidade das sequências, que decai rápidamente com o número de 1's presente na sequência. Por fim, temos o produto dos dois gráficos anteriores, representando assim a probabilidade total das sequências de um determinado tipo. Podemos observar, que o pico ocorre em sequências constituídas por 10% de 1's. Verificamos ainda que este pico torna-se mais acentuado com o crescimento de n. No limite, quando n for muito grande, só ocorrerão sequências com 10% de 1's.

Codificação para o Conjunto Típico

Após definirmos o que são sequências típicas e formalizarmos a propriedade da equipartição assintótica, vamos agora abordar o problema

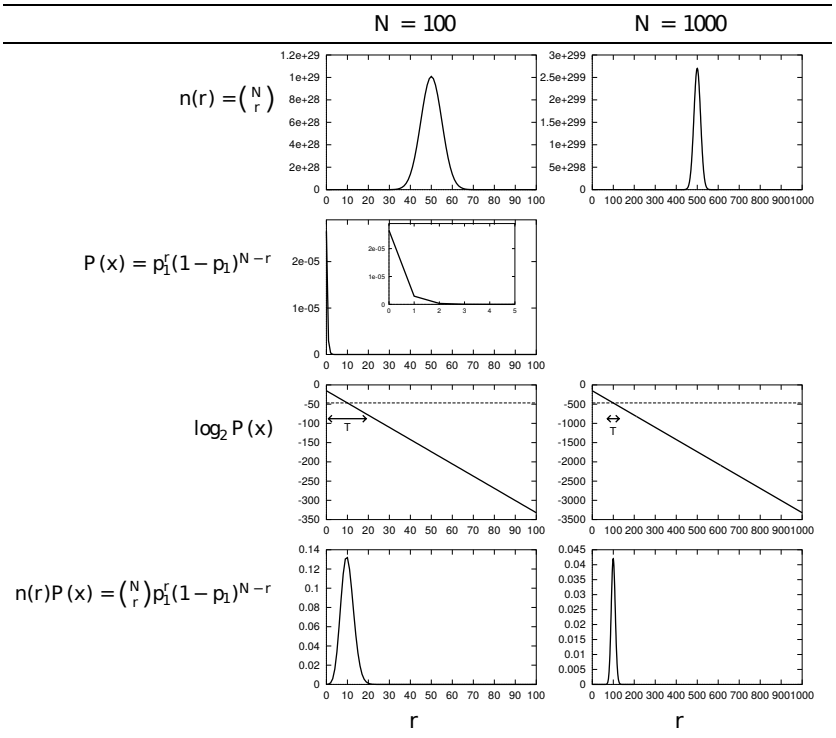

Figura 15: Para $p = 0.1$, $n = 100$ e $n = 1000$ os gráficos ilustram $n(r)$, o número de strings contendo r 1s; a probabilidade $P(x_{1:n})$ para uma string contendo r 1s; a mesma probabilidade em escala logarítmica; e a probabilidade total $n(r)P(x_{1:n})$ de todas as strings contendo r 1s (MacKay 2003).

de codificação considerando a codificação de sequências típicas. Como vimos que existe um limite para o tamanho do conjunto típicos, iremos utilizar o número de bits mínimo necessários para codificar as sequências quer pertençam a este conjunto. As sequências remancessentes, contidas no conjunto não típico, podem ser codificadas utilizando mais bits, sem que isto cause impacto significativo ao código, uma vez que a probabilidade do conjunto típico é aproximadamente 1 (conforme item 2 do Teorema 20).

A ideia consiste em particionar o conjunto de sequencias em dois blocos: conjunto típico $A_\epsilon^{(n)}$, e o seu complemento, o conjunto não típico $A_\epsilon^{(n)c} \triangleq \mathcal{X}^n \backslash A_\epsilon^{(n)}$. Tais partições satisfazem $A_\epsilon^{(n)} \cap A_\epsilon^{(n)c} = \emptyset$, e $A_\epsilon^{(n)} \cup A_\epsilon^{(n)c} = \mathcal{X}^n$. A Figura 16 ilustra o particionamento proposto.

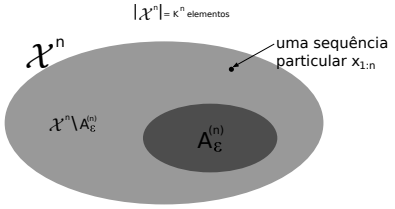

Figura 16: Particionamento de $\mathcal{X}^n$ em dois conjuntos: $A_\epsilon^{(n)}$ e $A_\epsilon^{(n)c}$.

Vamos então indexar os elementos de cada conjunto (conjunto típico e não-típico) separadamente. O número de elementos do conjunto típico é $|A_\epsilon^{(n)}| \leq 2^{n(H+\epsilon)}$, desta forma será necessário

$$\lceil n(H+\epsilon) \rceil \leq n(H+\epsilon) + 1\text{bits} \tag{183}$$

para indexar os elementos deste conjunto. Ainda, utilizaremos um bit extra para indicar se o elemento está no conjunto típico ou não, i.e., vamos utilizar uma sequência $(b_0, b_1, b_2, \ldots, b_{\lceil n(H+\epsilon) \rceil})$ onde o primeiro bit indica se temos um elemento do conjunto típico ($b_0 = 0$) ou não

e os demais indexam o elemento do conjunto. O número total de bits necessário para uma sequência típica será então $\leq n(H+\epsilon)+2$.

Para os elementos do conjunto não típico, vamos utilizar

$$\lceil \log |\mathcal{X}|^n \rceil \leq n \log K + 1 \text{bits}. \tag{184}$$

Então teremos um vetor binário da forma $(b_0, b_1, b_2, \ldots, b_{\lceil \log |\mathcal{X}|^n \rceil})$ onde $b_0 = 1$, indicando a atipicidade. Assim, o número total de bits para indexar uma sequência atípica será $\leq n \log K + 2$.

Este código proposto é 1-pra-1, sendo fácil codificar e decodificar, dado o *codebook*. Para o conjunto não-típico $A_\epsilon^{(n)c}$ estamos utilizando mais bits do que o necessário, já que $|A_\epsilon^{(n)c}| = |\mathcal{X}^n| - |A_\epsilon^{(n)}| = K^n - |A_\epsilon^{(n)}| \leq K^n$, mas isto não importará, como veremos adiante. O importante desta proposta de código é que as sequências típicas possuem um comprimento descritivo curto, aproximadamente nH.

Definição 18 (Comprimento da palavra associada a uma sequência)**.** O comprimento da palavra (*codeword*) associada à sequência $x_{1:n}$ é denominado $l(x_{1:n})$.

Assim, $l(X_{1:n})$ é uma variável aleatória, já que $X_{1:n}$ é uma variável aleatória. Então $El(X_{1:n}) = \sum_{x_{1:n}} p(x_{1:n}) l(x_{1:n})$ é o valor esperado do comprimento do código. Queremos que ele seja o menor possível.

Suponha que n seja grande suficiente de forma que $p(A_\epsilon^{(n)}) > 1-\epsilon$, então

$$\begin{aligned}
El(X_{1:n}) &= \sum_{x_{1:n}} p(x_{1:n}) l(x_{1:n}) && \text{(185a)}\\
&= \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) l(x_{1:n}) + \sum_{x_{1:n} \in A_\epsilon^{(n)c}} p(x_{1:n}) l(x_{1:n}) && \text{(185b)}\\
&\leq \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) [n(H+\epsilon)+2] + \sum_{x_{1:n} \in A_\epsilon^{(n)c}} p(x_{1:n}) [n \log K + 2] && \text{(185c)}\\
&= \underbrace{p(A_\epsilon^{(n)})}_{\leq 1} [n(H+\epsilon)+2] + \underbrace{p(A_\epsilon^{(n)c})}_{<\epsilon} [n \log K + 2] && \text{(185d)}\\
&\leq n(H+\epsilon) + 2 + \epsilon n \log K + \epsilon 2 && \text{(185e)}\\
&= n[H + \underbrace{\epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}}_{\epsilon'}] = n(H+\epsilon'), && \text{(185f)}
\end{aligned}$$

onde definimos ϵ' como

$$\epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}. \tag{186}$$

Podemos fazer ϵ' tão pequeno quanto queremos, fazendo ϵ pequeno e n grande. Desta forma, podemos fazer $n(H+\epsilon')$ tão próximo quanto quisermos de nH.

Teorema 21 (Primeiro Teorema de Shannon) *Seja $X_{1:n}$ i.i.d. $\sim p(x)$, $\epsilon > 0$, então $\exists$ um código $f_n : \mathcal{X}^n \to$ string binária e um inteiro n_ϵ, tal que o mapeamento seja um-pra-um (desta forma inversível sem erro), e*

$$E[\frac{1}{n} l(X_{1:n})] \leq H(X) + \epsilon \tag{187}$$

para todo $\epsilon > 0$ e todo $n \geq n_\epsilon$.

Demonstração. A demonstração do Primeiro Teorema de Shannon usa a codificação de sequências típicas de maneira semelhante ao exemplo anterior.

□

É interessante notar que a sequência mais provável não pertence ao conjunto típico $A_\epsilon^{(n)}$. Isso ocorre porque o conjunto típico contém sequências cujo histograma empírico é próximo da distribuição subjacente. A tipicidade não está relacionada à probabilidade de uma única sequência ser a maior.

Dado que o conjunto típcio é tal que $P(A_\epsilon^{(n)}) \to 1$ à medida que $n \to \infty$, ou seja, $A_\epsilon^{(n)}$ captura toda probabilidade, podemos nos questionar se existe um conjunto ainda menor que também contenha essencialmente 'toda' a probabilidade.

Definição 19 (Sequências asintoticamente iguais até a primeira ordem do expoente)**.** A notação $a_n \doteq b_n$ indica que a_n e b_n são iguais até a primeira ordem do expoente, ou seja, suas taxas de crescimento exponencial são as mesmas à medida que $n \to \infty$. Formalmente, isso significa que $\lim_{n\to\infty} \frac{1}{n} |\log a_n - \log b_n| = 0$, de modo que $\frac{1}{n} \log a_n \to \alpha$ e $\frac{1}{n} \log b_n \to \alpha$ para o mesmo α. Para n grande, a_n e b_n possuem aproximadamente o mesmo comportamento.

Teorema 22 *Seja $X_{1:n}$ uma sequência i.i.d. $\sim p(x)$. Para $\delta < 1/2$ e qualquer $\delta' > 0$, se $P(B_\delta^{(n)}) > 1 - \delta$, então*

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta', \tag{188}$$

se n é grande suficiente. Teremos assim

$$|B_\delta^{(n)}| > 2^{n(H-\delta')} \approx 2^{nH} \tag{189}$$

Usando a Definição 19, podemos então reescrever o teorema anterior da seguinte forma:

Se $\delta_n \to 0$ e $\epsilon_n \to 0$, então teremos

$$|B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH}. \tag{190}$$

Demonstração. Seja $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x)$. Seja $B_\delta^{(n)} \subset \mathcal{X}^n$ tal que $\Pr(B_\delta^{(n)}) > 1 - \delta$. Fixe $\epsilon < 1/2$. Dados dois subconjuntos quaisquer

A e B tais que $\Pr(A) > 1-\epsilon_1$ e $\Pr(B) > 1-\epsilon_2$. Seja A^c o complemento de A e B^c o complemento de B, então

$$P(A^c \cup B^c) \leq P(A^c) + P(B^c). \tag{191}$$

Como $P(A) > 1 - \epsilon_1$, teremos $P(A^c) \leq \epsilon_1$. De forma similar, $P(B^c) \leq \epsilon_2$. Poderemos assim escrever

$$\begin{aligned} P(A \cap B) &= 1 - P(A^c \cup B^c) && \text{(192a)} \\ &\geq 1 - P(A^c) - P(B^c) && \text{(192b)} \\ &\geq 1 - \epsilon_1 - \epsilon_2. && \text{(192c)} \end{aligned}$$

Podemos reescrever a desigualdade anterior como

$$\Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \geq 1 - \epsilon - \delta. \tag{193}$$

A probabilidade de um conjunto é dada pela soma das probabilidades de todos os elementos (sequências) neste conjunto, logo teremos

$$\Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) = \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \tag{194}$$

A probabilidade dos elementos no conjunto típico é limitada por $2^{-n(H-\epsilon)}$.

Desta forma teremos

$$\begin{aligned} 1 - \epsilon - \delta &\leq \Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) && \text{(195a)} \\ &= \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) && \text{(195b)} \\ &\leq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} && \text{(195c)} \\ &= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} && \text{(195d)} \\ &\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}, && \text{(195e)} \end{aligned}$$

onde utilizamos $A_\epsilon^{(n)} \cap B_\delta^{(n)} \subseteq B_\delta^{(n)}$.

Poderemos reescrever então da seguinte forma,

$$|B_\delta^{(n)}| > 2^{n(H-\epsilon)}, \tag{196}$$

onde $\epsilon > 0$.

□

Método de Tipos

O método de tipos refina a abordagem das sequências típicas, oferecendo uma análise mais estruturada das propriedades assintóticas. O método de tipos considera todas as possíveis distribuições empíricas

(ou 'tipos') que sequências $x_{1:n}$ podem ter, agrupando-as em conjuntos de sequências que compartilham um mesmo histograma empírico $\hat{p}_n(x_{1:n})$, ou $P_{x_{1:n}}$. O conjunto de sequências de comprimento n pode ser particionado em subconjuntos de sequências de tipos distintos. Na AEP, o método de tipos confirma que o número de sequências típicas é $\doteq 2^{nH(X)}$. Vamos então estabelecer algumas definições.

Definição 20 (Tipo)**.** Seja $X_1, X_2, \ldots, X_n \equiv X_{1:n}$ uma amostra de comprimento n de uma variável aleatória discreta D-ária. Então $x_i \in \mathcal{X}$ e o tamanho do alfabeto é $D = |\mathcal{X}|$, e $\mathcal{X} = \{a_1, a_2, \ldots, a_D\}$. Definimos a seguinte estatística, o histograma empírico da amostras, também chamado tipo da amostra:

$$P_{x_{1:n}} \triangleq \left(\frac{n(a_1|x_{1:n})}{n}, \frac{n(a_2|x_{1:n})}{n}, \ldots, \frac{n(a_D|x_{1:n})}{n} \right) \tag{197}$$

onde $n(a_i|x_{1:n})$ representa o número de ocorrências do símbolo a_i na amostra $x_{1:n}$.

Note que $P_{x_{1:n}}$ pode ser considerado como uma função massa probabilidade. Um tipo $\hat{p}$ é uma função de massa de probabilidade sobre S_X, definido como o histograma empírico $\hat{p}(a) = \frac{k_a}{n}$, onde $k_a = \sum_{i=1}^{n} \mathbb{I}\{x_i = a\}$ é o número de ocorrências do símbolo $a \in S_X$ na sequência, e $\sum_{a \in S_X} k_a = n$.

Definição 21 (Conjunto de Tipos)**.** O Conjunto de Tipos $\mathcal{P}_n$, ou $\mathcal{P}_n(\mathcal{X})$, para sequências de comprimento n sobre um alfabeto finito $\mathcal{X}$ é o conjunto de todas as possíveis distribuições empíricas (ou tipos) que podem ser formadas por sequências $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$.

Exemplo 9 (Ensaio de Bernoulli)**.** Para sequências de comprimento n geradas a partir de um ensaio de Bernoulli com alfabeto $\mathcal{X} = \{0, 1\}$, teremos o seguinte conjunto de tipos:

$$\mathcal{P}_n(\mathcal{X}) = \left\{ \left(\frac{0}{n}, \frac{n}{n} \right), \left(\frac{1}{n}, \frac{n-1}{n} \right), \ldots, \left(\frac{n}{n}, \frac{0}{n} \right) \right\} \tag{198}$$

onde existem $n + 1$ tipos (histogramas empíricos). O primeiro tipo em 198 representa a sequência sem nenhuma ocorrência de zeros e n ocorrências de uns; o segundo tipo representa as sequências com apenas uma ocorrência de zero e $n - 1$ ocorrência de uns; e o último tipo representa a sequência com n ocorrência de zeros e nenhuma ocorrência de uns.

Definição 22 (Classe de Tipo)**.** A Classe de Tipo $T(P)$, ou $T(\hat{p})$, associada a um tipo $\hat{p} \in \mathcal{P}_n(\mathcal{X})$, para sequências de comprimento n sobre um alfabeto finito $\mathcal{X}$, é o conjunto de todas as sequências $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ que compartilham um mesmo histograma empírico $\hat{p}$, ou P.

$$T(P) \triangleq \{x_{1:n} \in \mathcal{X}^n : P_{x_{1:n}} = P\}. \tag{199}$$

Exemplo 10 (Ensaio de Bernoulli)**.** Para o caso do ensaio de Bernoulli, vejamos alguns exemplos de classe de tipo.

Para o tipo $P = \left(\frac{0}{n}, \frac{n}{n}\right)$, a classe de tipo $T(P)$ possui apenas uma sequência:

$$T(P) = \{111\cdots11\}. \tag{200}$$

Para o tipo $P = \left(\frac{1}{n}, \frac{n-1}{n}\right)$, a classe de tipo $T(P)$ possui n sequências:

$$T(P) = \{011\cdots11, 101\cdots11, \ldots, 111\cdots01, 111\cdots10\}. \tag{201}$$

Para o tipo $P = \left(\frac{n}{n}, \frac{0}{n}\right)$, a classe de tipo $T(P)$ possui apenas uma sequência:

$$T(P) = \{000\cdots00\}. \tag{202}$$

Como mencionado anteriormente, o conjunto de todas sequências $\mathcal{X}^n$ pode ser particionado em diferentes conjuntos $T(P_i)$, com $P_i \in \mathcal{P}_n$, o conjunto de todos os tipos, ou seja, $\mathcal{P}_n = \{P_1, P_2, \ldots, P_{|\mathcal{P}_n|}\}$. As partições são disjuntas

$$T(P_i) \cap T(P_j) = \emptyset, \forall i, j \in \{1, \ldots, |\mathcal{P}_n|\} \text{ e } i \neq j, \tag{203}$$

e o particionamento é exaustivo

$$\bigcup_{P \in \mathcal{P}_n} T(P) = \mathcal{X}^n. \tag{204}$$

Este particionamento é ilustrado na Figura 17.

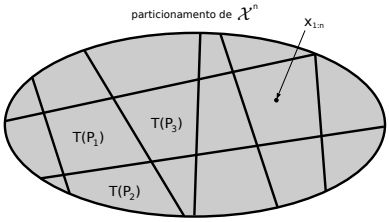

Figura 17: Particionamento do espaço de sequências de comprimento n.

No exemplo do ensaio de Bernoulli é fácil constatar quantos tipos distintos existem. O Teorema 24 a seguir usa o método estrela traço da análise combinatória para determinar o número de tipos existentes. Vejamos então primeiramente o teorema da análise combinatória.

Teorema 23 (Método Estrela Traço) *Suponha que n estrelas que devem ser organizadas em k recipientes, podendo inclusive ter recipiente vazio. Neste caso, o número de maneiras de fazer tal organização é dado por*

$$\binom{n+k-1}{k-1} \tag{205}$$

Demonstração. Este problema é o mesmo que incluir $k-1$ traços para separar uma sequência de n estrelas. Os traços podem ser inseridos em qualquer posição, inclusive sem existir estrelas entre eles, como por exemplo, com $n = 7$ e $k = 4$:

$$\star\star\star\star \;\;|\;\; |\;\; \star \;\; |\;\; \star\star$$

Note que, neste problema temos $n + k - 1$ símbolos (estrelas e traços) para serem incluídos, dos quais $k - 1$ são escolhidos para serem traços. Desta forma existem $\binom{n+k-1}{k-1}$ formas de dispor estes símbolos.

□

Teorema 24 (Número de tipos existentes) *Para sequências de comprimento n, em um alfabeto $\mathcal{X}$, o número de tipos é dado por*

$$|\mathcal{P}_n| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \tag{206}$$

Demonstração. Um histograma empírico pode ser visto como uma variação do problema estrela-traço. Se temos um alfabeto com $|\mathcal{X}| = D$ símbolos, o Conjunto de Tipos $\mathcal{P}_n$ contém todas as possíveis distribuições empíricas $\hat{p}$ para sequências de comprimento n sobre o alfabeto $\mathcal{X}$. Um tipo $\hat{p}$ é uma função de massa de probabilidade tal que $\hat{p}(a) = {}^{k_a}/_{n}$, onde k_a é o número de ocorrências do símbolo $a \in \mathcal{X}$, e $\sum_{a\in\mathcal{X}} k_a = n$. Assim, $k_a \in \{0, 1, \ldots, n\}$, e a restrição é que a soma de todos k_a's deve ser igual a n, o comprimento da sequência. Cada tipo $\hat{p}$ é definido pela ênupla $(k_{a_1}, k_{a_2}, \ldots, k_{a_D})$, e cada uma delas corresponde a uma forma de organização no problema de estrelas e traços. □

Embora tenhamos uma fórmula fechada para o número de tipos, Equação (206), iremos encontrar um limite superior para $|\mathcal{P}_n|$, que, embora seja um limite relativamente largo, é mais fácil de manipular e suficiente para muitas demonstrações teóricas.

Teorema 25 (Limite no número de tipos) *O número de tipos para sequências de comprimento n em um alfabeto $\mathcal{X}$ é limitado por*

$$|\mathcal{P}_n| \leq (n + 1)^{|\mathcal{X}|}. \tag{207}$$

Demonstração. O numerador de cada entrada em um tipo pode assumir $(n + 1)$ valores distintos (de 0 a n). Existem $|\mathcal{X}|$ entradas em um tipo, e portanto a mesma quantidade de numeradores. Os valores dos numeradores interagem entre si (a soma de todos deve ser igual a n), mas podemos achar um limite superior desconsiderando esta interação.

$$|\mathcal{P}_n| \leq \underbrace{(n + 1) \times (n + 1) \times \ldots \times (n + 1)}_{|\mathcal{X}| \text{ vezes}} = (n + 1)^{|\mathcal{X}|}. \tag{208}$$

□

É importante notar na Equação (207) que existe no máximo um número polinomial em n de tipos de sequencias de comprimento n. Entretanto, sabemos que o número de sequências cresce com n, $|\mathcal{X}|^n$. Com n grande suficiente, eventualmente, teremos que apenas um único tipo (o tipo das sequências típicas) conterá todas as sequências, como será demonstrado no Teorema 32.

Para sequências geradas por uma fonte i.i.d., a probabilidade de uma sequência $(x_1, x_2, \ldots, x_n)$ depende apenas do seu tipo e não da sequência específica. Isso significa que todas as sequências em uma mesma classe de tipo, ou seja, sequências que compartilham um mesmo histograma empírico, têm a mesma probabilidade.

Teorema 26 (Probabilidade Depende do Tipo) *Seja $X_1, X_2, \ldots, X_n$ i.i.d. $\sim Q(x)$, com Q arbitrário, e extensão $Q^n(x_{1:n}) = \prod_i Q(x_i)$, a probabilidade da sequencia depende apenas do tipo, ou seja, a probabilidade é 'independente' da sequencia, dado o tipo e Q, isto é,*

$$Q^n(x_{1:n}) = 2^{-n[H(P_{x_{1:n}})+D(P_{x_{1:n}}||Q)]}. \tag{209}$$

Demonstração.

$$\begin{aligned}
Q^n(x_{1:n}) &= \prod_{i=1}^{n} Q(x_i) = \prod_{a\in\mathcal{X}} Q(a)^{n(a|x_{1:n})} && \text{(210a)}\\
&= \prod_{a\in\mathcal{X}} Q(a)^{nP_{x_{1:n}}(a)} = \prod_{a\in\mathcal{X}} 2^{\left\{nP_{x_{1:n}}(a)\log Q(a)\right\}} && \text{(210b)}\\
&= \prod_{a\in\mathcal{X}} 2^{n\left\{P_{x_{1:n}}(a)\log Q(a) \underbrace{-P_{x_{1:n}}(a)\log P_{x_{1:n}}(a)+P_{x_{1:n}}(a)\log P_{x_{1:n}}(a)}_{=0}\right\}} && \text{(210c)}\\
&= 2^{n\sum_{a\in\mathcal{X}}\left(-P_{x_{1:n}}(a)\log\frac{P_{x_{1:n}}(a)}{Q(a)}+P_{x_{1:n}}(a)\log P_{x_{1:n}}(a)\right)} && \text{(210d)}\\
&= 2^{-n\left(D(P_{x_{1:n}}||Q)+H(P_{x_{1:n}})\right)}. && \text{(210e)}
\end{aligned}$$

□

Se Q é uma distribuição racional (i.e., um tipo possível) e se $x_{1:n} \in T(Q)$, então

$$Q^n(x_{1:n}) = 2^{-nH(Q)}. \tag{211}$$

Se Q for irracional, podemos fazer $D(P_{x_{1:n}} \,||\, Q)$ tão pequeno quando desejável, fazendo n grande suficiente.

No método de tipos, uma questão natural é identificar qual classe de tipo tem a maior probabilidade total quando as sequências são geradas por uma determinada distribuição. Intuitivamente, esperamos que a classe de tipo mais provável seja aquela cujo histograma empírico esteja mais próximo da distribuição geradora.

Lema 27 (Classe de Tipo com maior probabilidade) *Dada a distribuição geradora $Q = P \in \mathcal{P}_n$, teremos que $T(P)$ possui a maior probabilidade. Isto é*

$$P^n(T(P)) \geq P^n(T(\hat{P})),\ \forall \hat{P} \in \mathcal{P}_n. \tag{212}$$

Demonstração.

$$\frac{P^n(T(P))}{P^n(T(\hat{P}))} = \frac{|T(P)|\prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})|\prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \tag{213a}$$

$$= \frac{\binom{n}{nP(a_1)\ nP(a_2)\ \ldots\ nP(a_D)}\prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\hat{P}(a_1)\ n\hat{P}(a_2)\ \ldots\ n\hat{P}(a_D)}\prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \tag{213b}$$

$$= \prod_{a\in\mathcal{X}} \frac{[n\hat{P}(a)]!}{[nP(a)]!} P(a)^{n(P(a)-\hat{P}(a))} \tag{213c}$$

$$\geq \prod_{a\in\mathcal{X}} (nP(a))^{n(\hat{P}(a)-P(a))} P(a)^{n(P(a)-\hat{P}(a))} \tag{213d}$$

$$= \prod_{a\in\mathcal{X}} n^{n(\hat{P}(a)-P(a))} \tag{213e}$$

$$\geq \prod_{a\in\mathcal{X}} n^{n(\hat{P}(a)-P(a))} \tag{213f}$$

$$= n^{n\left[\sum_{a\in\mathcal{X}}\hat{P}(a)-\sum_{a\in\mathcal{X}}P(a)\right]} \tag{213g}$$

$$= n^{n(1-1)} = 1, \tag{213h}$$

onde em 213d utilizamos que $\frac{m!}{n!} \geq n^{m-n}$, para m e n inteiros não negativos[13]. Por fim, temos que $P^n(T(P)) \geq P^n(T(\hat{P}))$.

□

[13] Sejam m e n inteiros não negativos, então $\frac{m!}{n!} \geq n^{m-n}$. Se $m > n$, então $\frac{m!}{n!} = m(m-1)\ldots(n+1) \geq n^{m-n}$. Se $m < n$, então $\frac{m!}{n!} = \frac{1}{n(n-1)\ldots(m+1)} \geq \frac{1}{n^{n-m}}$. Se $m = n$, $\frac{m!}{n!} = 1 = n^0$.

O tamanho de uma classe de tipo $T(P)$ pode ser determinado pelos coeficientes multinomiais. Cada sequência $x_{1:n} \in T(P)$ possui exatamente $nP(a)$ ocorrências de cada símbolo $a \in \mathcal{X}$. O número de sequências corresponde ao número de maneiras de permutar os símbolos de acordo com o número de ocorrência de cada símbolo.

Lema 28 (Tamanho da Classe de Tipo) *Seja $P \in \mathcal{P}_n$, um tipo para sequências de comprimento n sobre um alfabeto finito $\mathcal{X}$, de tamanho $D = |\mathcal{X}|$, e seja $T(P)$ a classe de tipo associada, ou seja, o conjunto de todas as sequências $x_{1:n} \in \mathcal{X}^n$ com histograma empírico P. Então o tamanho de $T(P)$ é dado por*

$$|T(P)| = \binom{n}{nP(a_1)\ nP(a_2)\ \ldots\ nP(a_D)} = \frac{n!}{\prod_{a\in\mathcal{X}}(nP(a))!}. \tag{214}$$

Demonstração. Cada sequência em $T(P)$ tem exatamente $nP(a)$ ocorrências de cada símbolo $a \in \mathcal{X}$, com $\sum_{a\in\mathcal{X}}(nP(a)) = n$. O número de sequências distintas é o número de maneiras de permutar os n símbolos, considerando que as permutações dentro de cada símbolo não geram sequências distintas. Assim, o número de permutações distintas é o coeficiente multinomial, como dado na Equação (214).

□

Novamente, apesar de termos o resultado exato, dado pela Equação (214), em muitas situação é mais prático utilizarmos limites que

nos forneçam uma notação mais intuitiva e prática para utilização em outras demonstrações. Veremos então a seguir os limites superior e inferior para o tamanho da classe de tipo.

Teorema 29 (Limites no tamanho da Classe de Tipo) *Dado um tipo $P \in \mathcal{P}_n$, temos*

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}. \tag{215}$$

Demonstração (limite superior).

$$1 \geq P^n(T(P)) = \sum_{x_{1:n} \in T(P)} P^n(x_{1:n}) = \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \tag{216a}$$

$$= |T(P)| 2^{-nH(P)} \tag{216b}$$

□

Demonstração (limite inferior).

$$1 = \sum_{Q \in \mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q \in \mathcal{P}_n} \max_{R \in \mathcal{P}_n} P^n(T(R)) \tag{217a}$$

$$= \sum_{Q \in \mathcal{P}_n} P^n(T(P)) \leq (n+1)^{|\mathcal{X}|} P^n(T(P)) \tag{217b}$$

$$= (n+1)^{|\mathcal{X}|} \sum_{x_{1:n} \in T(P)} P^n(x_{1:n}) \tag{217c}$$

$$= (n+1)^{|\mathcal{X}|} \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \tag{217d}$$

$$\leq (n+1)^{|\mathcal{X}|} \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \tag{217e}$$

$$= (n+1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)}, \tag{217f}$$

onde em 217b utilizamos

$$P = \operatorname*{argmax}_{R \in \mathcal{P}_n} P^n(T(R)). \tag{218}$$

Ao final, obtemos

$$|T(P)| \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)}. \tag{219}$$

□

Para o caso binário, $\mathcal{X} = \{0, 1\}$, o Teorema 29 nos fornece os seguintes seguintes limites para o tamanho da classe de tipo:

$$\frac{1}{(n+1)^2} 2^{nH(\frac{k}{n})} \leq \binom{n}{k} \leq 2^{nH(\frac{k}{n})}, \tag{220}$$

entretanto, o limite inferior pode ser ainda mais restrito, fornecendo assim

$$\frac{1}{(n+1)}2^{nH(\frac{k}{n})} \leq \binom{n}{k} \leq 2^{nH(\frac{k}{n})}. \tag{221}$$

Os limites superior e inferior para o tamanho de uma classe de tipo $T(P)$, dados no Teorema 29, podem ser utilizados para derivar limites correspondentes na probabilidade total da classe de tipo $Q^n(T(P))$, onde Q é a distribuição subjacente da fonte i.i.d. que gera as sequências. Um resultado importante é que qualquer outro tipo que seja menos próximo de Q (em termos de divergência de Kullback-Leibler) terá sua probabilidade decrescendo exponencialmente com n, decrescendo assim mais rapidamente que o tipo mais provável.

Teorema 30 (Limites da probabilidade da classe de tipo) *Para qualquer $P \in \mathcal{P}_n$ e seja Q a distribuição subjacente da fonte i.i.d., a probabilidade da classe de tipo $T(P)$ sob Q^n é tal que $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$. Especificamente, temos os limites*

$$\frac{1}{(n+1)^{|\mathcal{X}|}}2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)}. \tag{222}$$

Demonstração.

$$Q^n(T(P)) = \sum_{x_{1:n}\in T(P)} Q^n(x_{1:n}) = \sum_{x_{1:n}\in T(P)} 2^{-n(D(P||Q)+H(P))} \tag{223a}$$

$$= |T(P)|2^{-n(D(P||Q)+H(P))} \tag{223b}$$

para completar a demonstração, os limites do tamanho da classe de tipo dados pela Equação (215).

□

Iremos agora rever a definição de conjunto típico, utilizando a formalização do métodos de tipos.

Definição 23 (Conjunto Típico)**.** Seja $X_1, X_2, \ldots, X_n$ i.i.d. $\forall i$, $X_i \sim Q(x)$. Então o conjunto típico é definido como

$$T_Q^\epsilon = \{x_{1:n} : D(P_{x_{1:n}} \,||\, Q) \leq \epsilon\}. \tag{224}$$

Podemos agora analizar a probabilidade do conjunto típico.

Teorema 31 (Probabilidade do Conjunto Típico) *Sejam $X_1, X_2, \ldots, X_n$ i.i.d. $\forall i$, $X_i \sim Q(x)$. A probabilidade do complemento do conjunto típico $\overline{T}_Q^\epsilon$ é dada por*

$$Q(\overline{T}_Q^\epsilon) = Q(\{x_{1:n} : D(P_{x_{1:n}} \,||\, Q) > \epsilon\}) \leq 2^{-n(\epsilon - |\mathcal{X}|\frac{\log(n+1)}{n})}. \tag{225}$$

Dessa forma,

$$D(P_{x_{1:n}} \,||\, Q) \xrightarrow{p} 0, \text{ quando } n \to \infty, \tag{226}$$

ou seja, a divergência entre $P_{x_{1:n}}$ e Q converge em probabilidade para zero quando n é grande suficiente.

Demonstração.

$$Q(\overline{T}_Q^{\epsilon}) = \sum_{P\in\mathcal{P}_n: D(P||Q)>\epsilon} Q^n(T(P)) \tag{227a}$$

$$\leq \sum_{P\in\mathcal{P}_n: D(P||Q)>\epsilon} 2^{-nD(P||Q)} \tag{227b}$$

$$\leq \sum_{P\in\mathcal{P}_n: D(P||Q)>\epsilon} 2^{-n\epsilon} \tag{227c}$$

$$\leq (n+1)^{|\mathcal{X}|}2^{-n\epsilon} = 2^{-n\left(\epsilon-|\mathcal{X}|\frac{\log(n+1)}{n}\right)} \tag{227d}$$

então a probabilidade vai para zero quando $n \to \infty$, e desta forma a probabilidade do conjunto típico vai para 1 quando $n \to \infty$.

□

Os tipos que divergem (KL) mais do que ϵ da distribuição subjacente Q terão probabilidade decrescente. Para n grande, o conjunto típico acaba sendo a única coisa que ocorre com uma probabilidade não evanescente.

Dizemos que uma codificação de fonte é universal quando ela não depende da distribuição fonte. Seria possível criar um código universal que atinja o limite da entropia, ou seja, uma taxa $R > H(Q)$ (em bits por símbolo), onde $H(Q)$ é a entropia da distribuição subjacente Q? O método de tipos oferece uma abordagem para formalizar o teorema de Shannon, permitindo uma análise mais refinada das sequências com base em seus histogramas empíricos. Existem no máximo $2^{nH(P)}$ sequências do tipo P. Podermos utilizar $nH(P)$ bits para representar tais sequencias. Se $R > H(P)$, podemos utilizar nR bits para representar estas sequências. Quando n cresce, apenas os tipos P 'próximos' de Q irão ocorrer.

A Figura 18 ilustra um codificador de blocos que associa cada uma das M sequências de n de símbolos da fonte a uma sequência binária de comprimento m. Neste codificadador de blocos, cada sequências de n símbolos é codificada conjuntamente associando a ela uma palavra a cada. O codificador faz o mapeamento de sequências de tamanho n produzidas pela fonte em sequências de m bits. M é o número de possíveis mensagens e também o número de palavras do *codebook* do codificador.

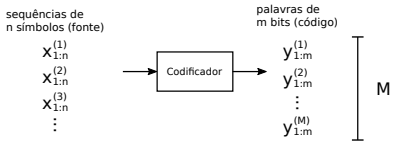

Figura 18: Codificador (M, n), onde M representa o número de mensagens e n o comprimento das sequências.

Definição 24 (Código de Bloco com Taxa Fixa R)**.** Seja $X_1, X_2, \ldots, X_n \sim Q$, i.i.d. mas Q desconhecido. A função do codificador e decodificador são definidas a seguir:

$$\text{codificador: } f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\} \tag{228}$$

$$\text{decodificador: } \phi_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n \tag{229}$$

e a probabilidade de erro

$$P_e^{(n)} = Q^n(\{x_{1:n} : \phi(f_n(x_{1:n})) \neq x_{1:n}\}). \tag{230}$$

Definição 25 (Código de Bloco Universal de Taxa R)**.** Um código de bloco de taxa R para um fonte é dito universal se a função f_n e ϕ_n não depender da distribuição Q e se

$$P_e^{(n)} \to 0 \text{ quando } n \to \infty \text{ sempre que } H(Q) < R. \tag{231}$$

Veremos que, se $R > H(Q)$, então existe uma sequência (em n) de códigos com erro evanescente. Por outro lado, se $R < H(Q)$ a probabilidade de erro vai pra 1.

Na demonstração do Teorema 32 utilizaremos ainda a definição de simplex probabilístico.

Definição 26 (Simplex Probabilístico)**.** O Simplex Probabilístico em $\mathbb{R}^m$ é o conjunto de pontos $x_{1:m} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ tal que $x_i \geq 0$, $\sum_{i=1}^m x_i = 1$.

Exemplo 11 (Simplex Probabilístico com $m = 2$)**.** O Simplex probabilístico será o conjunto de pontos

$$\{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 = 1\}, \tag{232}$$

representado na Figura 19.

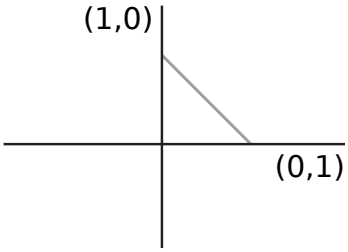

Figura 19: Simplex probabilístico para $m = 2$.

Exemplo 12 (Simplex Probabilístico com $m = 3$)**.** O Simplex probabilístico será o conjunto de pontos

$$\{(x_1, x_2, x_3) : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_1 + x_2 + x_3 = 1\} \tag{233}$$

representado na Figura 20 e com alguns pontos em destaque na Figura 21.

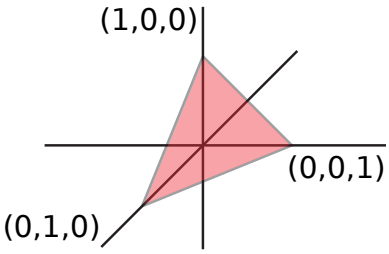

Figura 20: Simplex probabilístico para $m = 3$.

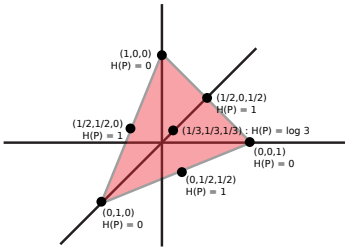

Figura 21: Representação de alguns pontos e suas respectivas entropias.

O Teorema da Codificação de Shannon é um dos resultados fundamentais da teoria da informação, estabelecendo as condições sob as quais é possível representar as informações produzidas por uma fonte i.i.d. com distribuição subjacente Q sem que haja perdas, utilizando, para tanto uma taxa R. O Teorema mostra que, se $R > H(Q)$, então existe um código de bloco universal para representar a informação produzida pela fonte com probabilidade de erro tendendo a zero à medida que o comprimento n das sequências aumenta. Por outro lado, se $R < H(Q)$, não é possível representar a informação da fonte com probabilidade de erro evanescente. No contexto do método de tipos, podemos formalizar esse teorema explorando a estrutura das classes de tipo: como o número de tipos é polinomial em n, enquanto o número de sequências é exponencial em n, e apenas os tipos próximos de Q ocorrem, para n grande, é possível construir códigos universais que codifiquem eficientemente as sequências produzidas pela fonte, usando aproximadamente $nH(Q)$ bits.

Teorema 32 (Teorema da Codificação de Shannon) $\exists$ *uma sequencia* $(2^{nR}, n)$ *de códigos universais tais que* $P_e^{(n)} \to 0$ *para toda distribuição* Q *tal que* $H(Q) < R$.

Demonstração. Fixe $R > H(Q)$. Defina uma taxa para n que é fixada a um fator polinomial:

$$R_n \triangleq R - |\mathcal{X}|\frac{\log(n+1)}{n} < R. \tag{234}$$

Defina um conjunto de sequências que possuem entropia de tipo menor do que esta taxa:

$$A_n \triangleq \{x_{1:n} \in \mathcal{X}^n : H(P_{x_{1:n}}) \leq R_n\} \tag{235a}$$

$$= \left\{ \bigcup_{P \in \mathcal{P}_n} T(P) : H(P) \leq R_n \right\} \tag{235b}$$

A partir desta definição, teremos que

$$|A_n| = \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} |T(P)| \leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nH(P)} \tag{236a}$$

$$\leq \sum_{P \in \mathcal{P}_n : H(P) \leq R_n} 2^{nR_n} \leq (n+1)^{|\mathcal{X}|} 2^{nR_n} \tag{236b}$$

$$= 2^{n\left(R_n + |\mathcal{X}|\frac{\log(n+1)}{n}\right)} = 2^{nR}. \tag{236c}$$

Como $|A_n| \leq 2^{nR}$, podemos indexar A_n com nR bits.

O codificador será dado por

$$f_n(x_{1:n}) = \begin{cases} \text{índice de } x_{1:n} \text{ em } A_n, & \text{se } x_{1:n} \in A_n \\ 0, & \text{caso contrário .} \end{cases} \tag{237}$$

O codificador associará um índice a $x_{1:n}$ se $H(P_{x_{1:n}}) \leq R_n$ (ou seja, $x_{1:n} \in A_n$); e não associará valor se $H(P_{x_{1:n}}) > R_n$ (ou seja, $x_{1:n} \notin A_n$). Note que $f_n(\cdot)$ não depende da distribuição da fonte, apenas do ordenamento e de $\mathbb{R}^m$. Um erro ocorrerá se $x_{1:n} \notin A_n$. Os tipos podem ser representados por pontos em um Simplex Probabilístico, conforme ilustrado na Figura 22.

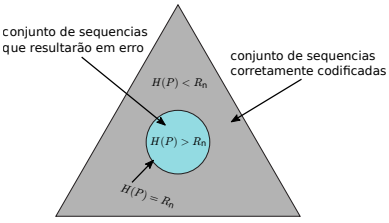

Figura 22: Representação das sequências em um simplex probabilísticos. As sequências com $H(P_{x_{1:n}}) \leq R_n$ não acarretarão em erro no processo de codificação.

Um erro ocorre quando a sequência não está em A_n. Desta forma,

$$P_e^{(n)} = 1 - Q^n(A_n) = \quad Q^n(A_n^c) = \sum_{P : H(P) > R_n} Q^n(T(P)) \tag{238a}$$

$$\leq \quad \sum_{P : H(P) > R_n} \max_{P : H(P) > R_n} Q^n(T(P)) \tag{238b}$$

$$\leq \quad (n+1)^{|\mathcal{X}|} \max_{P : H(P) > R_n} Q^n(T(P)) \tag{238c}$$

$$\leq \quad (n+1)^{|\mathcal{X}|} \max_{P : H(P) > R_n} 2^{-nD(P||Q)} \tag{238d}$$

$$= \quad (n+1)^{|\mathcal{X}|} 2^{-n[\min_{P : H(P) > R_n} D(P||Q)]} \tag{238e}$$

onde utilizamos que $Q^n(T(P)) \leq 2^{-nD(P||Q)}$. Temos que R_n forma uma sequência crescente com n, tal que $R_n < R$ para todo n (veja a Figura 23). Por hipótese, $H(Q) < R$. Eventualmente, para algum n_0, teremos que $\forall n > n_0$, $R_n > H(Q)$. Na Equação (238e), escolhemos $P : H(P) > R_n$.

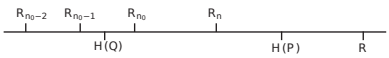

Figura 23: Sequência de taxas R_n.

Teremos então: $H(P) > R_n > H(Q)$, o que implica em $P \neq Q$. Desta forma, teremos $D(P \mid\mid Q) > 0$, para P que foi escolhido em 238e. Teremos assim

$$P_e^{(n)} \leq \underbrace{(n+1)^{|\mathcal{X}|}}_{\text{polinomial em } n} \underbrace{2^{-n[\min_{P:H(P)>R_n} D(P||Q)]}}_{\text{exp. decrescente qnd } n\to\infty} \tag{239}$$

Logo, $P_e^{(n)} \to 0$ quando $n \to \infty$.

□

Por outro lado, se $R < H(Q)$ teremos $P_e^{(n)} \to 1$. A entropia é então o limite de representação, ou compressão.

Bibliografia

Araújo, Leonardo Carneiro, Thaïs Cristófaro-Silva e Hani Camille Yehia (2013). "Entropy of a Zipfian Distributed Lexicon". Em: *Glottometrics* 26, pp. 38–49.

Berger, Roger L. e George Casella (2002). *Statistical Inference*. Duxbury Press.

Cover, Thomas M. e Joy A. Thomas (2006). *Elements of Information Theory*. 2ª ed. Hoboken, NJ: Wiley-Interscience. ISBN: 978-0-471-24195-9.

Delfs, Hans e Helmut Knebl (2015). *Introduction to Cryptography: Principles and Applications*. Springer Berlin Heidelberg. ISBN: 9783662479742. DOI: 10.1007/978-3-662-47974-2.

Ferrer i Cancho, Ramon e Ricard V. Solé (dez. de 2001). "Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited∗". Em: *Journal of Quantitative Linguistics* 8.3, pp. 165–173. ISSN: 1744-5035. DOI: 10.1076/jqul.8.3.165.4101. URL: http://dx.doi.org/10.1076/jqul.8.3.165.4101.

Gallager, Robert Gray (jan. de 1962). "Low-density parity-check codes". Em: *IEEE Transactions on Information Theory* 8.1, pp. 21–28. ISSN: 0018-9448. DOI: 10.1109/tit.1962.1057683.

Golay, Marcel JE (1949). "Notes on digital coding". Em: *Proc. IEEE* 37, p. 657.

Hamming, Richard Wesley (abr. de 1950). "Error Detecting and Error Correcting Codes". Em: *Bell System Technical Journal* 29.2, pp. 147–160. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1950.tb00463.x.

Hartley, Ralph Vinton Lyon (jul. de 1928). "Transmission of Information". Em: *Bell System Technical Journal* 7.3, pp. 535–563. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1928.tb01236.x.

MacKay, David John Cameron (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK ; New York: Cambridge University Press. ISBN: 9780521642989.

Shannon, Claude E. (1948). "A mathematical theory of communication." Em: *Bell Syst. Tech. J.* 27.3, pp. 379–423.

Yeung, Raymond W. (2002). *A First Course in Information Theory*. Springer US. ISBN: 9780306467912.

Zipf, George Kingsley (1935). *An Introduction to Dynamic Philology*. Cambridge, MA: The MIT Press.

– (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.