Teoria da Informação

Prof. Leonardo Araújo

https://sites.google.com/site/leolca/teaching/information-theory

https://github.com/leolca/lectures/blob/master/ti/teoria-da-informacao.pdf

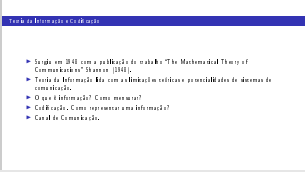
- Regra da Cadeia
- Translação e Mudança de Escala
- Desigualdade de Hadamard
- Entropia Máxima e Gaussianas
- Erro de Estimação

13 Canais Contínuos

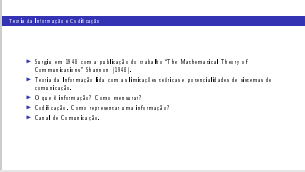
- Canal Gaussiano
- Canais em Paralelo

Teoria da Informação e Codificação

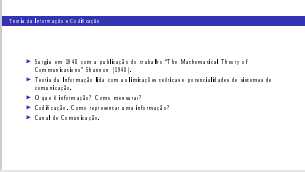
- Surgiu em 1948 com a publicação do trabalho "The Mathematical Theory of Communications" Shannon (1948).
- Teoria da Informação lida com as limitações teóricas e potencialidades de sistemas de comunicação.
- O que é informação? Como mensurar?
- Codificação. Como representar uma informação?
- Canal de Comunicação.



- A distinguibilidade entre as mensagens é fator importante para caracterizar informação. Pela definição de Shannon, "Informação é a habilidade de distinguir de forma confiável dentre um rol de alternativas possíveis".
- Shannon cunhou o termo 'auto-informação' de um evento ou mensagem aleatória definindo-o como "menos o logaritmo da probabilidade do evento aleatório". A 'entropia' da fonte estocástica que gera os eventos é o valor esperado da auto-informação.
- Shannon mostrou que a entropia de uma fonte estocástica possui um significado físico: em média, é o menor número de bits necessários para representar ou comunicar de forma fidedigna eventos gerados por uma fonte estocástica.
- O problema central em Teoria da Informação é a transmissão eficiente e confiável de dados, do transmissor a um receptor, através de um canal de comunicação.
- Eficiência (usar o mínimo de recursos possível).
- Confiabilidade (evitar erros, ser capaz de detectá-los e corrigi-los).

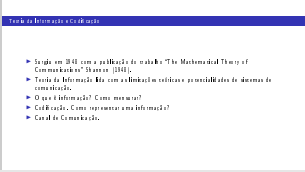


- A informação é um conceito paradoxal. Por um lado necessita de uma representação física, por outro lado é abstrata. Uma mesma informação pode ser representada em papel, em um meio magnético ou ótico, pode ser representada por ondas mecânicas ou elétricas.
- Linguagem. Comunicação falada e escrita. Código faz associação entre símbolo e mensagem e é 'arbitrário'.



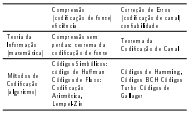
A área da ciência criada por Shannon ampliou-se ao longo do tempo e influência diversas outras áreas. Por exemplo: teoria da comunicação, criptografia, ciência da computação, física (mecânica estatística), matemática (probabilidade e estatística), filosofia da ciência, linguística e processamento de linguagem natural, reconhecimento de fala, reconhecimento de padrões e aprendizado de máquina, compressão de dados, economia, biologia e genética, psicologia, etc.

Shannon entrou para o Bell Labs para trabalhar com sistemas de controle de disparo e criptografia durante a Segunda Guerra Mundial, sob um contrato com o Comitê Nacional de Pesquisa para Defesa Em 1945, Shannon elaborou um memorando sigiloso, que posteriormente foi publicado sob o título "Communication Theory of Secrecy Systems". Este incorporava muitos dos conceitos e formulações matemáticas do artigo mais consagrado "A Mathematical Theory of Communication" (1948).



- Codificação - criar códigos com algoritmos práticos para codificação e decodificação para serem utilizados na comunicação no mundo real em canais ruidosos.
- Exemplos de codificações conhecidas para representar informação: código Morse, código ASCII, etc.

	Compressão (codificação de fonte) eficiência	Correção de Erros (codificação de canal) confiabilidade
Teoria da Informação (matemática)	Compressão sem perdas: teorema da codificação de fonte	Teorema da Codificação de Canal
Métodos de Codificação (algoritmo)	Códigos Simbólicos: código de Huffman Códigos de Fluxo: Codificação Aritmética, Lempel-Ziv	Códigos de Hamming, Códigos BCH Códigos Turbo Códigos de Gallager



	Compressão (codificação de fonte) sem ruído	Correcção de Erros (codificação do canal) com falibilidade
Teoria da Informação (probabilística)	Compressão sem perdas: teorema da codificação de fonte	Teorema da Codificação de Canal
Métodos de Codificação (algoritmos)	Códigos Similares: código de Huffman Códigos de Fluxo: Codificação Aritmética, Lempel-Ziv	Códigos de Hamming, Códigos BCH Códigos Turbo Códigos de Gallager

Huffman (1952) sem perdas; ótimo (dentre os códigos de símbolos); simples implementação; PNG, JPEG, MPEG, WinZip, GZip, MP3, AAC.

Codificação Aritmética (1978, patente IBM) Peter Elias (1963): trabalho não publicado; problema: precisão infinita (solucionado, Jorma Rissanen e Richard Pasco, 1976) patentes (IBM, todas já expiraram); JPEG, JBIG, Skype, PPM, PAQ, DjVu.

Lempel-Ziv (1978), Lempel–Ziv–Welch (1984) assintoticamente ótimo; eficiente e simples de ser implementado; PNG, GIF, PKZip, GZipg, PDF.

Códigos de Hamming verificação de paridade; primeiro código de correção de erros efetivamente bom (1950); RAM (podem ocorrer erros no processo de escrita e leitura), RAID 2.

Código de Reed-Solomon (1960) muitas vezes combinados com códigos convolucionais, por exemplo: RSV (algoritmo de vitterbi); robusto a erros em rajada; códigos de barra, CD, DVD, Blue-Ray, DSL (telefone), DVB; (Digital Video Broadcasting), RAID 6, comunicação satélite e sondas espaciais (Voyager 1977, Galileo 1989, Cassini-Huygens; 1997, Pathfinder 1996, MER 2003).

	Compressão (codificação de fonte) eficácia	Correção de Erros (codificação do canal) confiabilidade
Teoria da Informação (paramétrica)	Compressão sem perdas: teorema da codificação de fonte	Teorema da Codificação de Canal
Métodos de Codificação (algoritmos)	Códigos Símbolos: código de Huffman Código de Fano: Codificação Aritmética, Lempel-Ziv	Códigos de Hamming, Códigos BCH Códigos Turbo Códigos de Gallager

códigos Turbo (1993) 3G, 4G, LTE, WiMax, Mars Reconnaissance Orbiter (MRO) 2005.

códigos Gallager: low-density parity-check (LDPC) Robert G. Gallager, MIT (1960), não eram práticos para os computadores da época; David J. C. MacKay e Radford M. Neal (1996); 10Gb/s Ethernet, WiFi 802.11 N, Internet por rede elétrica ITU-T G.hn, DVB-S2 (tv via satélite), China Mobile Multimedia Broadcasting (CMMB) transmissão multimídia via satélite para aparelhos móveis.

Código Morse

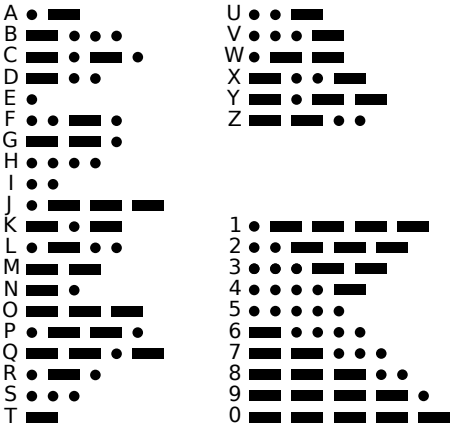


Figura 1: Código Morse internacional (Wikipedia (2020d)). Letras do alfabeto ordenadas por frequência de ocorrência no inglês: etaoin shrdlu cmfwyp vbgkjq xz (Wikipedia (2020a,c)).

Código Unário

Claude Mendibil utilizou o código unário (1, 01, 001, 0001, ...) para representar as letras do alfabeto ESARINTULOMDPCFBVHGJQZYXKW. Utilizando este código, Jean-Dominique Bauby ditou o livro *Le Scaphandre et le Papillon* (O Escafandro e a Borboleta).

Figura 2: Foto de Bauby em 1996 ditando suas memórias para Claude Mendibil (Wikipedia (2020b)).

Compressão

- Compressão é importante para utilizarmos melhor os recursos disponíveis.
- Shannon mostrou que o limite para a representação é a entropia.

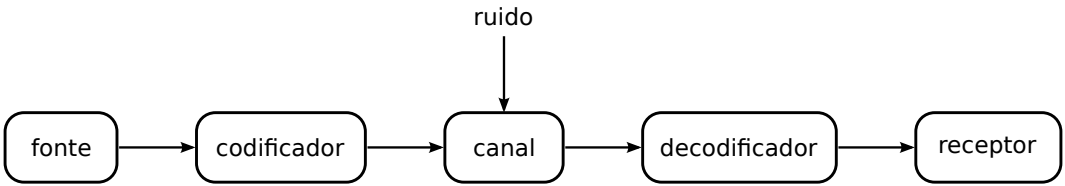
Compressão

- Compressão é importante para utilizarmos melhor os recursos disponíveis.
- Shannon mostrou que o limite para a representação é a entropia.

Compressão é importante para utilizarmos melhor os recursos disponíveis.

1. Armazenar mais dados em meio (disco rígido, memória, fita, etc).
2. Transmitir mais informação através de um canal (essencialmente, armazenar e transmitir são o mesmo problema).
3. Diminuir o desgaste do meio ao reduzir o número de vez que se faz leitura e escrita. Solid State Drives (SSDs) baseados em memórias flash NAND possuem um número finito de ciclos de programar/apagar. É importante reduzir a quantidade de bits que serão gravados para aumentar a vida útil dessas memórias/discos. (A compressão LZ4 vem sendo utilizada com esta finalidade, e também para que o S.O. tenha um boot mais rápido)

Modelo Geral de Comunicação



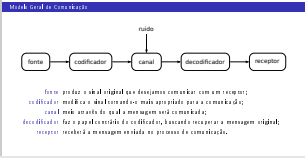
fonte produz o sinal original que desejamos comunicar com um receptor;

codificador modifica o sinal tornando-o mais apropriado para a comunicação;

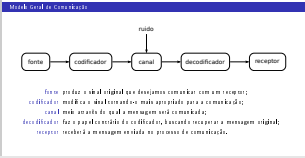
canal meio através do qual a mensagem será comunicada;

decodificador faz o papel contrário do codificador, buscando recuperar a mensagem original;

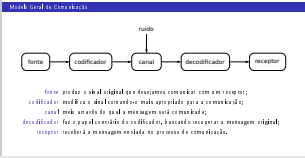
receptor receberá a mensagem enviada no processo de comunicação.



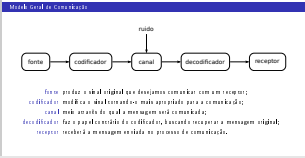
- Separação do codificador/decodificador em duas partes: codificador/decodificador de fonte e codificador/decodificador de canal.
- Remover redundância do sinal produzido pela fonte e acrescentar redundância por causa do ruído no canal de comunicação.
- Fontes e Canais de comunicação: discretos ou contínuos.



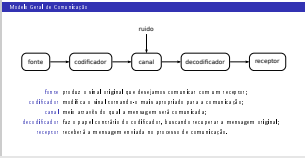
No contexto de Teoria da Informação, fonte é qualquer coisa que produza uma mensagem, um sinal que carregue informação. Podemos considerar uma fonte que produz mensagens como: voz, sons, palavras, imagens, vídeo, sequência de bits de um programa de computador, etc.



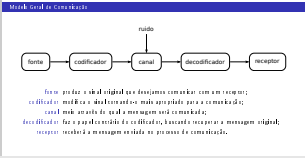
Canal é o meio através do qual o sinal produzido pela fonte será transmitido/propagado/armazenado. Por exemplo: espaço aberto (ar), linha telefônica, link de rádio, link em uma comunicação espacial, disco rígido, CD, DVD, fita magnética (armazenamento - transmissão no tempo ao invés de espaço pode sofrer deterioração ao longo do tempo); DNA de seres vivos ao longo de gerações, envio de mensagens por estímulos elétricos ou químicos em um organismo biológico.



Receptor é aquele a quem é destinada a mensagem transmitida. Exemplos: computador ou equipamento, uma pessoa, rádio, tv, sistema de áudio, etc.

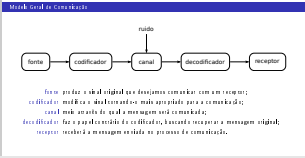


Ruído é qualquer sinal que interfere com aquele que está sendo transmitido. Exemplos: ruído térmico, ruído impulsivo, cross-talk, outro sinal qualquer indesejado. Ruído representa a nossa compreensão imperfeita do universo. Desta forma, tratamos ruído como algo aleatório e que usualmente obedece certas regras, tais como uma determinada distribuição probabilística.



O codificador processa o sinal antes de inseri-lo no canal de comunicação.

- Redução dos dados, removendo redundância do sinal.
- Inserção de redundâncias de acordo com as características do canal de comunicação, para garantir integridade aos dados transmitidos.
- Codificação para representar as informações de um sinal sob a forma de outro sinal.



O decodificador faz o papel inverso do codificador.

- Remove os erros de transmissão.
- Recupera a informação original enviada pela fonte.

Introdução - sugestões I

Khan Academy - information theory

https://www.khanacademy.org/computing/computer-science/informationtheory

Introdução - sugestões II

The Shannon Centennial: 1100100 years of bits
https://www.youtube.com/watch?v=pHSRHi17RKM

Claude Shannon - Father of the Information Age
https://www.youtube.com/watch?v=z2Whj_nL-x8

The Shannon Limit - Bell Labs - Future Impossible
https://www.youtube.com/watch?v=HSoog0OqgV0

Notação

X é uma variável aleatória

x é uma valor que a v.a. assume

$\mathcal{X}$ é o alfabeto de tamanho $|\mathcal{X}| = K$ dentro do qual a v.a. assume valores, $\mathcal{X} = \{a_1, \ldots, a_K\}$

$\mathcal{P}_X$ é o conjunto de probabilidades associadas aos valores, $\mathcal{P}_X = \{p_1, \ldots, p_K\}$, tais que $p_i \geq 0$ e $\sum_{i=1}^{K} p_i = 1$

p_i é a probabilidade da v.a. assumir um determinado valor, $p_i = \Pr(X = a_i)$

$\mathcal{P}_X = p$ é a distribuição da v.a., $\sum_{x \in \mathcal{X}} \Pr(X = x) = 1$

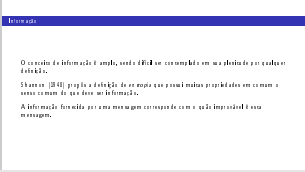
$X \sim p$, a v.a. X possui distribuição p

Informação

O conceito de informação é amplo, sendo difícil ser contemplado em sua plenitude por qualquer definição.

Shannon (1948) propôs a definição de *entropia* que possui muitas propriedades em comum o senso comum do que deve ser informação.

A informação fornecida por uma mensagem corresponde com o quão improvável é esta mensagem.



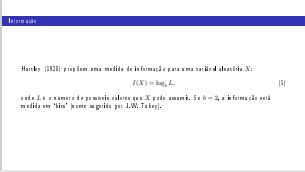
O que é previsível fornece pouca ou nenhuma informação.
Quanto mais incerto, mais informação há.

Informação

Hartley (1928) propõem uma medida de informação para uma variável aleatória X:

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

onde L é o número de possíveis valores que X pode assumir. Se $b = 2$, a informação será medida em 'bits' (nome sugerido por J.W. Tukey).



A definição de Hartley é condizente com as seguintes intuições sobre informação:

- Dois cartões de memória devem possuir o dobro da capacidade de um cartão para armazenamento de informação.
- Dois canais de comunicação idênticos devem possuir o dobro da capacidade de transmitir informação que um único canal.
- Um dispositivo com duas posições estáveis, como um relé ou um flip-flop, armazena um bit de informação. N dispositivos deste tipo podem armazenar N bits de informação, já que o número total de estados é 2^N e $\log_2 2^N = N$.

Entretanto, isto é válido apenas quando as mensagens/eventos são equiprováveis. No caso extremo, note que se o cartão de memória armazena apenas zeros, ele não é capaz de armazenar informação alguma.

Entropia

Suponha que existam eventos E_k com probabilidade de ocorrência p_k.

- Shannon: informação associada ao evento E_k é dada por $I(E_k) = \log(1/p_k)$.
  - Se $p_k = 1 \rightarrow$ não há surpresa na ocorrência do evento E_k.
  - Se $p_k = 0 \rightarrow$ surpresa infinita, afinal o evento E_k é impossível.
  - $I(E_k) = -\log p(E_k)$ é a auto-informação do evento ou mensagem E_k.
- **Sempre** utilizaremos a base 2 para o cálculo do logaritmo, desta forma $\log \equiv \log_2$, a menos que seja especificado o contrário.
- $\ln$ é o logaritmo na base natural e.

Entropia

- Notação: $p(x) = P_X(X = x)$, a probabilidade do evento $\{X = x\}$, da v.a. X assumir o valor x.
- Valor esperado da v.a. X: $E[X] = EX = \sum_x xp(x)$.
- Dada uma função $g : \mathcal{X} \to \mathbb{R}$, o valor esperado da v.a. $g(X)$ é $Eg(X) = \sum_x g(x)p(x)$.
- Considere $g(x) = \log(1/p(x))$. Então $g(x)$ é a imprevisão (surpresa) de encontrar o evento $X = x$. Tomando o valor esperado de g teremos

$$\sum_x p(x) \log \frac{1}{p(x)}, \tag{2}$$

ou seja, a esperança da surpresa, ou o valor esperado da imprevisão na variável aleatória X. Esta é a definição de entropia.

Entropia

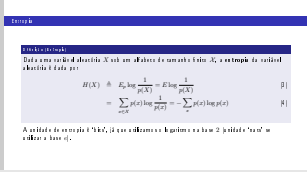
Definição (Entropia)

Dada uma variável aleatória X sob um alfabeto de tamanho finito $\mathcal{X}$, a **entropia** da variável aleatória é dada por

$$H(X) \triangleq E_p \log \frac{1}{p(X)} = E \log \frac{1}{p(X)} \tag{3}$$

$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = -\sum_{x} p(x) \log p(x) \tag{4}$$

A unidade de entropia é 'bits', já que utilizamos o logaritmo na base 2 (unidade 'nats' se utilizar a base e).



- Entropia mede a grau de incerteza associado a uma distribuição.
- Entropia mede a desordem ou o espalhamento de uma distribuição.
- Entropia mede a 'escolha' que a fonte tem na escolha de símbolos de acordo com uma densidade (maior entropia implica em mais escolha).
- Vamos utilizar a seguinte convenção: $0 \log 0 = 0$.

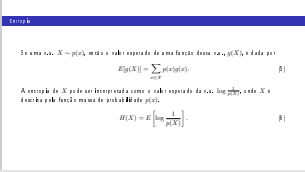
Entropia

Se uma v.a. $X \sim p(x)$, então o valor esperado de uma função desta v.a., $g(X)$, é dada por

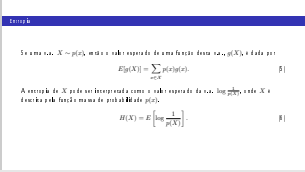
$$E[g(X)] = \sum_{x \in \mathcal{X}} p(x)g(x). \tag{5}$$

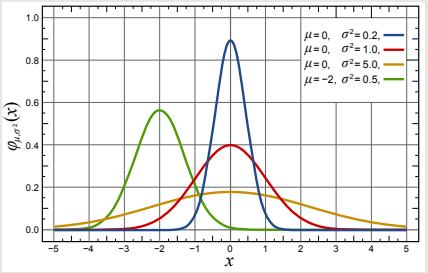
A entropia de X pode ser interpretada como o valor esperado da v.a. $\log \frac{1}{p(X)}$, onde X é descrita pela função massa de probabilidade $p(x)$.

$$H(X) = E\left[\log \frac{1}{p(X)}\right]. \tag{6}$$



- Entropia é uma medida da real 'incerteza' média, o que é uma medida sobre toda a distribuição.
- Entropia mede o grau de incerteza médio ou esperado do resultado de uma distribuição de probabilidade.
- É uma medida de desordem ou espalhamento. Distribuições com alta entropia devem ser planas, mais uniformes, enquanto distribuições com baixa entropia devem possuir poucas modas (unimodal, bimodal).





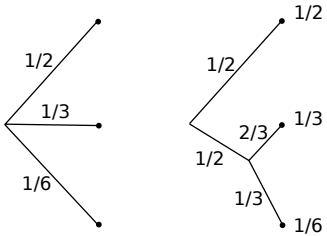
- mais concentrado: menor entropia
- mais espalhado: maior entropia
- os valores em x não importam, apenas os valores das probabilidades associadas $p(x)$ importam no cálculo da entropia

Escolha, Incerteza e Entropia I

Suponha um conjunto de eventos cujas probabilidades de ocorrências sejam dadas por $p_1, p_2, \ldots, p_n$. É possível encontrar uma medida de quanta 'escolha' está envolvida na seleção de um evento ou quão incertos estamos da saída?
Para tal medida $H(p_1, p_2, \ldots, p_n)$, é razoável requerermos as seguintes propriedades:

1) H deve ser contínuo em p_i;
2) Se todos os p_i são iguais, $p_i = \frac{1}{n}$, então H deve ser uma função monotonicamente crescente de n (quando temos eventos equiprováveis, teremos mais incerteza quão maior for o número de eventos possíveis);
3) Se for possível quebrar uma escolha em uma sequência de escolhas sucessivas, a medida H original deve ser a soma ponderada dos valores individuais das medidas H_i após a quebra.

Escolha, Incerteza e Entropia II



$$H(\frac{1}{2},\frac{1}{3},\frac{1}{6}) = H(\frac{1}{2},\frac{1}{2}) + \frac{1}{2}H(\frac{2}{3},\frac{1}{3}) \tag{7}$$

A única função H que satisfaz às suposições acima é da forma Shannon (1948):

$$H = -K\sum_{i=1}^{k} p(i)\log p(i)\ , \tag{8}$$

onde K é uma constante positiva.

Demonstração da Equação (8) I

Nesta secção iremos apresentar a demonstração de $H = -\sum p_i \log p_i$ (conforme Apêndice 2 de Shannon (1948)).
Vamos definir

$$A(n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right). \tag{9}$$

Desejamos 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.

Demonstração da Equação (8) II

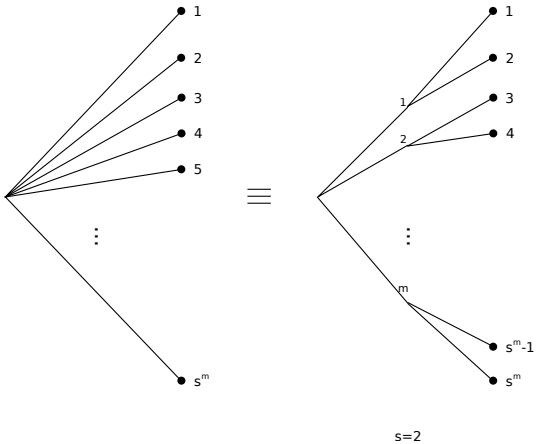


Figura 3: Exemplo de equivalência para $s = 2$.

Demonstração da Equação (8) III

Teremos então que

$$A(s^m) = mA(s). \tag{10}$$

Da mesma forma, para t e n, teremos $A(t^n) = nA(t)$. Podemos tomar n arbitrariamente grande e encontrar m que satisfaça

$$s^m \leq t^n \leq s^{(m+1)}. \tag{11}$$

Tomando o logaritmo[1] da expressão acima e dividindo por $n \log s$ todos os termos[2], teremos

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}, \tag{12}$$

o que é equivalente a

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \epsilon, \tag{13}$$

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

Demonstração da Equação (8) IV

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

$$\begin{aligned} A(s^m) \leq A(t^n) &\leq A(s^{(m+1)}) \\ mA(s) \leq nA(t) &\leq (m+1)A(s) \end{aligned} \tag{14}$$

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

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n}, \tag{15}$$

ou, de forma equivalente,

$$\left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \epsilon, \tag{16}$$

e assim, como as duas frações ($\log t / \log s$ e $A(t)/A(s)$) estão ϵ próximas de m/n, podemos concluir que

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| < 2\epsilon. \tag{17}$$

Demonstração da Equação (8) V

Como ϵ é arbitrariamente pequeno, no limite teremos

$$\begin{aligned} \frac{A(t)}{A(s)} &= \frac{\log t}{\log s} \\ A(t) &= \frac{A(s)}{\log s} \log t = K \log t, \end{aligned} \tag{18}$$

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, 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

Demonstração da Equação (8) VI

i-ésima dada escolha, realizar uma nova escolha dentre n_i opções igualmente prováveis. Teremos então:

$$\begin{aligned} \overbrace{K \log\left(\sum n_i\right)}^{A(\sum n_i)} &= H(p_1,\ldots,p_n) + \overbrace{K \log n_i}^{A(n_i)} \\ K \underbrace{\left(\sum p_i\right)}_{=1} \log\left(\sum n_i\right) &= H(p_1,\ldots,p_n) + K \underbrace{\left(\sum p_i\right)}_{=1} \log n_i. \end{aligned} \tag{19}$$

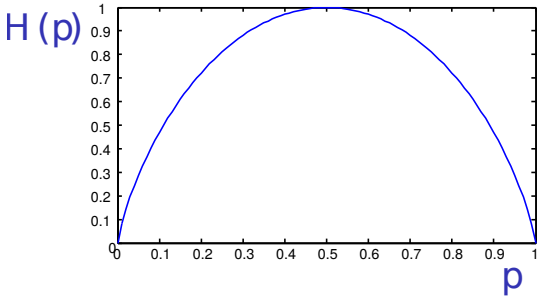
E assim,

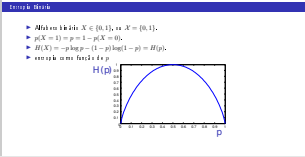
$$\begin{aligned} H(p_1,\ldots,p_n) &= K\left[\left(\sum p_i\right)\log\left(\sum n_i\right) - \left(\sum p_i\right)\log n_i\right] \\ &= -K\sum p_i \log\frac{n_i}{\sum n_i} = -K\sum p_i \log p_i. \quad \square \end{aligned} \tag{20}$$

[1] Logaritmo é uma função monótona crescente.
[2] $n \log s$ é positivo para $n \geq 0$ e $s \geq 1$.

Entropia Binária

- Alfabeto binário $X \in \{0, 1\}$, ou $\mathcal{X} = \{0, 1\}$.
- $p(X = 1) = p = 1 - p(X = 0)$.
- $H(X) = -p \log p - (1 - p) \log(1 - p) = H(p)$.
- entropia como função de p





- maior incerteza ($H = 1$) quando $p = 0.5$ e menor incerteza ($H = 0$) quando $p = 0$ ou $p = 1$.
- note que a entropia $H(p)$ é concava em p.

Entropia - GNU Octave

```
function H = entropy(p,b)

  if (nargin == 0 || nargin > 2) print_usage (); endif;
  if any(p < 0) | any(p > 1) | abs(sum(p)-1) > 1E-10, error('not a
      ↪ valid pmf!'); endif;

  id = find(p!=0);
  p = p(id);
  H = sum( - p .* log2(p) );

  if nargin > 1, H *= log(2)/log(b); endif;

endfunction
```

[download do código]

Entropia - GNU Octave - demo

```
%!demo
%! p = [0.5 0.5];
%! H = entropy(p);
%! printf('The pmf p has entropy = %.2f bits.\n',H);
%! He = entropy(p,e);
%! printf('The pmf p has entropy = %.2f nats.\n',He);
%! p = [0:0.02:1];
%! for i=1:length(p), H(i) = entropy([p(i), (1-p(i))]); endfor;
%! figure; plot(p,H); xlabel('p'); ylabel('H(p) (bits)'); title('
   ↪ binary entropy');
```

Entropia - Exemplo

Suponha uma v.a. $X \in \mathcal{X} = \{a, b, c, d\}$ com distribuição dada por

$$X = \begin{cases} a, & \text{com probabilidade } \frac{1}{2}, \\ b, & \text{com probabilidade } \frac{1}{4}, \\ c, & \text{com probabilidade } \frac{1}{8}, \\ d, & \text{com probabilidade } \frac{1}{8}. \end{cases} \tag{21}$$

A entropia associada será dada por

$$\begin{aligned} H(X) &= H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} \\ &= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = \frac{14}{8} = \frac{7}{4} \end{aligned} \tag{22}$$

Entropia Conjunta

Duas variáveis aleatórias X e Y possuem **entropia conjunta**

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = E\log\frac{1}{p(X,Y)}. \tag{23}$$

Generalizando para vetores $X_{1:N} = (X_1, X_2, \ldots, X_N)$

$$\begin{aligned} H(X_{1:N}) &= H(X_1, X_2, \ldots, X_N) \\ &= \sum_{x_1,x_2,\ldots,x_N} p(x_1,\ldots,x_N)\log\frac{1}{p(x_1,\ldots,x_N)} \\ &= E\log\frac{1}{p(X_1,\ldots,X_N)} \end{aligned} \tag{24}$$

Entropia Condicional I

Dadas duas v.a. X e Y relacionadas por $p(x, y)$, conhecer o evento $X = x$ pode alterar a entropia de Y.

- Entropia condicionada a um evento $H(Y|X = x)$

$$\begin{aligned} H(Y|X=x) &= E \log \frac{1}{p(Y|X=x)} \\ &= -\sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \end{aligned} \tag{25}$$

- $H(Y|X = x)$ é uma função de X. Podemos então tomar seu valor esperado $E[H(Y|X = x)]$ e obter a entropia condicional $H(Y|X)$.

Entropia Condicional II

- Realizando a média sobre todos os x, obteremos a entropia condicional $H(Y|X)$.

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= -\sum_{x,y} p(x,y) \log p(y|x) \\ &= E \log \frac{1}{p(Y|X)} \end{aligned} \tag{26}$$

Regra da Cadeia

Teorema (Regra da Cadeia para a Entropia)

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{27}$$

Demonstração.

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

tomando o valor esperado de ambos os lados, obtemos o resultado desejado. □

Corolário

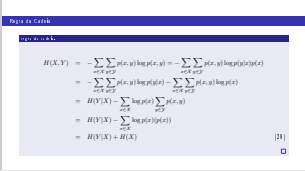
Se $X \perp\!\!\!\perp Y$ *então* $H(X,Y) = H(X) + H(Y)$.

Regra da Cadeia

regra da cadeia.

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)p(x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}} \log p(x)\sum_{y\in\mathcal{Y}} p(x,y) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}} \log p(x)(p(x)) \\
&= H(Y|X) + H(X) \qquad (29)
\end{aligned}
$$

□


Regra da Cadeia

Regra da cadeia.

$$\begin{aligned} H(X,Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)p(x) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) \\ &= H(Y|X) - \sum_{x\in X}\log p(x)\sum_{y\in Y} p(x,y) \\ &= H(Y|X) - \sum_{x\in X}\log p(x)p(x)) \\ &= H(Y|X) + H(X) \end{aligned}$$ |23|

Exemplo Canal de Comunicação

Suponha um canal de comunicação com entrada X e saída Y.
$H(X|Y)$ pode ser visto como a incerteza sobre X (a mensagem enviada) quando Y (a mensagem recebida) for conhecido.
Sem nenhuma observação no processo de comunicação através deste canal, o receptor não sabe nada sobre X nem Y, assim a incerteza inicial é $H(X,Y)$. Quando o receptor recebe a mensagem Y, ele ganha uma quantidade de informação $H(Y)$. Assim a informação que falta sobre X mesmo conhecendo Y é dada por $H(X|Y) = H(X,Y) - H(Y)$. Esta pode ser tido como uma medida do erro na comunicação.
A quantidade de informação que o receptor de fato ganha é $I(X;Y) = H(X) - H(X,Y)$.

Regra da Cadeia Generalizada I

Teorema (Regra da Cadeia para a Entropia)

$$H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, X_2, \ldots, X_{i-1}) \tag{30}$$

$$\begin{aligned} H(X_1, X_2, \ldots, X_N) &= H(X_1) + H(X_2|X_1) + \\ & H(X_3|X_1, X_2) + H(X_4|X_1, X_2, X_3) + \cdots \end{aligned} \tag{31}$$

Regra da Cadeia Generalizada II

Demonstração.

Utilizando a regra da cadeia da probabilidade condicional, teremos

$$p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1}), \tag{32}$$

então

$$-\log p(x_1, x_2, \ldots, x_N) = -\sum_{i=1}^{N} \log p(x_i | x_1, x_2, \ldots, x_{i-1}) \tag{33}$$

tomando o valor esperado de ambos os lados, obtemos o resultado desejado. □

Propriedades da Entropia

1) H é uma função estritamente côncava de X, i.e., para $0 \leq \lambda \leq 1$ e variáveis aleatórias X e Y

$$H(\lambda X + (1-\lambda)Y) \geq \lambda H(X) + (1-\lambda)H(Y) \tag{34}$$

com igualdade sse (se e somente se) $\lambda = 0$ ou $\lambda = 1$ ou $X = Y$.

2) $H(X) \geq 0$ com igualdade sse $p(X)$ for não nulo apenas em um ponto $x_0 \in \mathcal{X}$.
3) $H(X) \leq \log |\mathcal{X}|$ com igualdade sse $p(X)$ for uniforme ($p \sim \frac{1}{n}$).
4) $H(X)$ é uma função apenas das probabilidades $p(x_i)$, independente da ordem ou rótulo.
5) $H_b(X) = (\log_b a) H_a(X)$.

Continuidade da Entropia I

Todas as medidas de informação de Shannon são funções contínuas das distribuições conjuntas das variáveis aleatórias envolvidas.

Definição (distância das variações)

Seja p e q duas distribuições probabilísticas em um alfabeto comum $\mathcal{X}$. A distância das variações entre p e q é definida por

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|. \tag{35}$$

Dado um alfabeto finito fixo $\mathcal{X}$, considere $\mathcal{P}_{\mathcal{X}}$ o conjunto de todas as distribuições em $\mathcal{X}$. A entropia para uma dada distribuição p sobre o alfabeto $\mathcal{X}$ é definida por

$$H(p) = -\sum_{x \in S_p} p(x) \log p(x), \tag{36}$$

Continuidade da Entropia II

onde S_p denota o suporte de p, ou seja, $S_p \subset \mathcal{X}$.
Para que $H(p)$ seja contínuo com respeito à convergência em distância das variações, em uma determinada distribuição $p \in \mathcal{P}_\mathcal{X}$, devemos ter que, para qualquer $\epsilon > 0$, existe $\delta > 0$ tal que

$$|H(p) - H(q)| < \epsilon, \tag{37}$$

para todo $q \in \mathcal{P}_\mathcal{X}$ satisfazendo

$$V(p, q) < \delta, \tag{38}$$

ou, de forma equivalente,

$$\lim_{p' \to p} H(p') = H\left(\lim_{p' \to p} p'\right) = H(p), \tag{39}$$

onde a convergência $p' \to p$ é em distância das variações.

Continuidade da Entropia III

Como $a \log a \to 0$ quando $a \to 0$, definimos uma função $l : [0, \infty) \to \mathbb{R}$ da forma

$$l(a) = \begin{cases} a \log a & \text{se } a > 0, \\ 0 & \text{se } a = 0, \end{cases} \tag{40}$$

ou seja, $l(a)$ é uma extensão contínua de $a \log a$. Podemos reescrever a entropia da seguinte forma

$$H(p) = -\sum_{x \in \mathcal{X}} l(p(x)), \tag{41}$$

onde o somatório é tomado em todo $x \in \mathcal{X}$ ao invés de S_p. Definindo uma função $l_x : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$, para todo $x \in \mathcal{X}$, da forma

$$l_x(p) = l(p(x)), \tag{42}$$

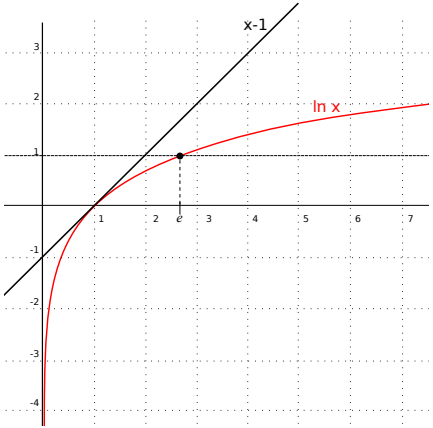
teremos

$$H(p) = -\sum_{x \in \mathcal{X}} l_x(p). \tag{43}$$

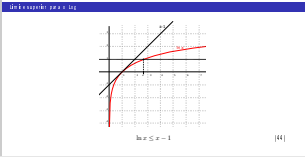
Continuidade da Entropia IV

Evidentemente $l_x(p)$ é contínua em p (com relação à convergência em distância das variações). Como o somatório na Equação 43 possui apenas um número finito de termos, podemos concluir que $H(p)$ é uma função contínua de p.

Limite superior para o Log



$$\ln x \leq x - 1 \tag{44}$$



$\ln x \leq x - 1$, para $x \geq 1$

- Sabemos que para $x = 1$ é verdadeiro, $0 = \ln 1 \leq 1 - 1 = 0$.
- Vamos demonstrar que $\ln x \leq x - 1$, para $x \geq 1$ por contradição.

Suponha que existe $b > 1$ tal que $\ln x > x - 1$, para $x = b$. Vamos definir $f(x) = \ln x - x + 1$, 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

$$f'(c) = \frac{f(b) - f(1)}{b - 1} = \frac{\overbrace{f(b)}^{>0}}{\underbrace{b - 1}_{>0}} > 0. \tag{45}$$

Mas, $f'(x) = 1/x - 1$, e assim $f'(x) < 0$ para $x > 1$. Logo há uma contradição e nossa hipótese é falsa. Teremos assim $\ln x \leq x - 1$, para $x \geq 1$.
Para $x \in (0, 1)$, basta seguir os mesmos passos, escolhendo um ponto $x = b$, 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.

Valor Máximo da Entropia (discreta) I

Teorema (Limite Superior da Entropia)

Seja $X \in \{x_1, x_2, \ldots, x_n\}$. *Então* $H(X) \leq \log n$, *sendo a igualdade alcançada se e somente se* $p(X = x_i) = \frac{1}{n}$ *para todo* i.

Valor Máximo da Entropia (discreta) II

Demonstração.

Vamos mostrar que $H(X) - \log n \leq 0$.

$$\begin{aligned} H(X) - \log n &= -\sum_x p(x) \log p(x) - \log n \overbrace{\sum_x p(x)}^{=1} \\ &= -\sum_x p(x) \log p(x) - \sum_x p(x) \log n \\ &= -\sum_x p(x) \log p(x) n = \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n} \\ &\leq \log_2 e \sum_x p(x) \left[\frac{1}{p(x)n} - 1 \right] \end{aligned}$$

...

Valor Máximo da Entropia (discreta) III

Demonstração.

continuação...

$$
\begin{aligned}
H(X) - \log n &\leq \dots \\
&= \log_2 e \left[\underbrace{\sum_x \frac{1}{n}}_{\sum_{x \in \mathcal{X}} \frac{1}{n} = n \frac{1}{n} = 1} \underbrace{- \sum_x p(x)}_{=1} \right] = 0 \qquad (46)
\end{aligned}
$$

□

Valor Máximo da Entropia

Na demonstração acima utilizamos $\ln z \leq z - 1$. A igualdade $\ln z = z - 1$ se dará no ponto estacionário $z = 1$, isto é, quando $\frac{1}{p(x)n} = 1$, ou seja, quando $p(x) = 1/n$, teremos assim uma distribuição uniforme.

Se tivermos $p_i = 1/n$, então

$$-\sum_i p_i \log p_i = -\sum_i \frac{1}{n} \log \frac{1}{n} = -\log \frac{1}{n} = \log n. \tag{47}$$

Podemos mostrar (através da concavidade da entropia) que este é o único conjunto de valores com esta propriedade.

- Entropia aumenta quando a distribuição se torna mais uniforme.

Valor Máximo da Entropia I

Outra demonstração...

Demonstração.

Considere $X \in \mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ com probabilidades $p = \{p_1, p_2, \ldots, p_n\}$, respectivamente. A entropia de X é dada por

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p_i \log p_i \\ &= -\sum_{i=1}^{n-1} p_i \log p_i - p_n \log p_n \\ &= -\left(\frac{1}{\ln 2}\right)\left[\sum_{i=1}^{n-1} p_i \ln p_i + p_n \ln p_n\right]. \end{aligned} \tag{48}$$

...

Valor Máximo da Entropia II

Demonstração.

continuação...

Da mesma forma, podemos expressa p_n da seguinte maneira

$$p_n = 1 - \sum_{i=1}^{n-1} p_i. \tag{49}$$

Utilizando 49 em 48, podemos expressar a entropia $H(X)$ como uma função de $n-1$ probabilidades p_i. O máximo será dado quando a seguinte condição ocorrer

$$\frac{\partial H(X)}{\partial p_k} = 0 \quad \text{for} \quad k = 1, \ldots, n-1 \ . \tag{50}$$

...

Valor Máximo da Entropia III

Demonstração.

continuação...

Teremos então

$$
\begin{aligned}
0 = \frac{\partial H(X)}{\partial p_k} &= -\left(\frac{1}{\ln 2}\right)\frac{\partial}{\partial p_k}\left[\sum_{i=1}^{n-1} p_i \ln p_i + p_n \ln p_n\right] \\
&= -\left(\frac{1}{\ln 2}\right)\left[\ln p_k + 1 + (\ln p_n + 1)\frac{\partial p_n}{\partial p_k}\right] \\
&= -\left(\frac{1}{\ln 2}\right)\left[\ln p_k + 1 - (\ln p_n + 1)\right], \qquad (51)
\end{aligned}
$$

onde utilizamos a Equação 49, que nos fornece $\partial p_n / \partial p_k = -1$.

...

Valor Máximo da Entropia IV

Demonstração.

continuação...

A Equação 51 mostra que devemos encontrar $\ln p_k = \ln p_n$ para cada $k = 1, \ldots, n-1$. Todas as $n-1$ equações serão satisfeitas quando todas as probabilidades p_k forem iguais a $1/n$.

...

Valor Máximo da Entropia V

Demonstração.

continuação...

Devemos agora calcular a derivada segunda para mostrar que o extremo que achamos é de fato um máximo.

$$\begin{aligned} \frac{\partial^2 H(X)}{\partial p_k^2} &= -\left(\frac{1}{\ln 2}\right)\frac{\partial}{\partial p_k}\left[\ln p_k - \ln p_n\right] \\ &= -\left(\frac{1}{\ln 2}\right)\left[\frac{1}{p_k} + \frac{1}{p_n}\right] \leq 0 \ , \end{aligned} \tag{52}$$

já que as probabilidades são valores positivos. Escolhendo então $p_k = 1/n$, teremos a entropia máxima.

□

Subdividindo a entropia em partes I

A entropia deve permanecer inalterada, mesmo quando subdividimos as escolhas em partes.

Exemplo (Exemplo simples)

Suponha uma v.a. X com alfabeto $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ e distribuição $q = (q_1.q_2, q_3, q_4)$.
A entropia associada a esta variável aleatória é dada por

$$\begin{aligned} H(X) &= H(q_1, q_2, q_3, q_4) \\ &= -\sum_{i=1}^{4} q_i \log q_i \\ &= -q_1 \log q_1 - q_2 \log q_2 - q_3 \log q_3 - q_4 \log q_4. \end{aligned} \tag{53}$$

...

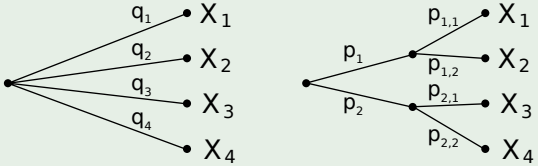
Subdividindo a entropia em partes II

Exemplo (Exemplo simples)

continuação...

Se dividirmos a escolha na determinação de X em duas escolhas sucessivas, conforme ilustrado na figura abaixo, poderemos então escrever

$$H(q_1, q_2, q_3, q_4) = H(p_1, p_2) + p_1 H(p_{1,1}, p_{1,2}) + p_2 H(p_{2,1}, p_{2,2}). \tag{54}$$



...

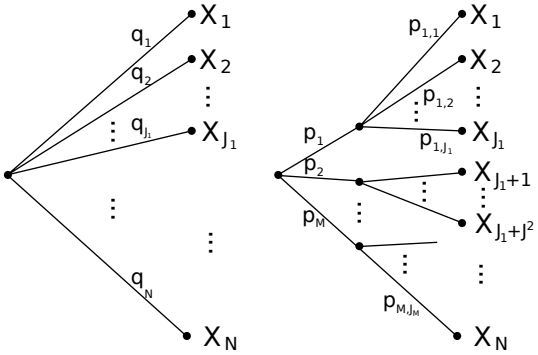
Subdividindo a entropia em partes III

Exemplo (Exemplo simples)

continuação...

$$
\begin{aligned}
H(X) &= -q_1 \log q_1 - q_2 \log q_2 - q_3 \log q_3 - q_4 \log q_4 \\
&= -p_1 p_{1,1} \log p_1 p_{1,1} - p_1 p_{1,2} \log p_1 p_{1,2} - p_2 p_{2,1} \log p_2 p_{2,1} - p_2 p_{2,2} \log p_2 p_{2,2} \\
&= -p_1 p_{1,1} \log p_1 - p_1 p_{1,1} \log p_{1,1} - p_1 p_{1,2} \log p_1 - p_1 p_{1,2} \log p_{1,2} \ldots \\
&\quad -p_2 p_{2,1} \log p_2 - p_2 p_{2,1} \log p_{2,1} - p_2 p_{2,2} \log p_2 - p_2 p_{2,2} \log p_{2,2} \\
&= -p_1 \log p_1 \left(p_{1,1} + p_{1,2}\right) - p_2 \log p_2 \left(p_{2,1} + p_{2,2}\right) \ldots \\
&\quad +p_1 \left(-p_{1,1} \log p_{1,1} - p_{1,2} \log p_{1,2}\right) \ldots \\
&\quad +p_2 \left(-p_{2,1} \log p_{2,1} - p_{2,2} \log p_{2,2}\right) \\
&= H(p_1, p_2) + p_1 H(p_{1,1}, p_{1,2}) + p_2 H(p_{2,1}, p_{2,2}) \qquad (55)
\end{aligned}
$$

Subdividindo a entropia em partes IV



Subdividindo a entropia em partes V

De forma geral, como $q_n = p_m p_{m,j}$, teremos

$$\begin{aligned}
H(X) &= -\sum_{n=1}^{N} q_n \log q_n = -\sum_{m=1}^{M}\sum_{j=1}^{J_m} p_m p_{m,j} \log p_m p_{m,j} \\
&= -\sum_{m=1}^{M}\sum_{j=1}^{J_m} (p_m p_{m,j} \log p_m + p_m p_{m,j} \log p_{m,j}) \\
&= -\sum_{m=1}^{M}\sum_{j=1}^{J_m} p_m p_{m,j} \log p_m - \sum_{m=1}^{M}\sum_{j=1}^{J_m} p_m p_{m,j} \log p_{m,j} \\
&= -\sum_{m=1}^{M} p_m \log p_m \left(\sum_{j=1}^{J_m} p_{m,j}\right) - \sum_{m=1}^{M} p_m \sum_{j=1}^{J_m} p_{m,j} \log p_{m,j} \\
&= H(p_1, \ldots, p_M) + \sum_{m=1}^{M} p_m H(p_{m,1}, \ldots, p_{m,J_m}). \qquad (56)
\end{aligned}$$

Embaralhar

- Suponha que X seja uma v.a. indicando as posições de cartas (i.e. $X = x$ representa um conjunto de posições, uma determinada configuração).
- Seja T uma operação de embaralhamento independente, i.e. $T \perp\!\!\!\perp X$.
- Então $H(TX) \geq H(X)$.

$$
\begin{aligned}
H(TX) &\geq H(TX|T) \\
&\quad \text{onde utilizamos que condicionar não pode aumentar a entropia, como veremos adiante} \\
&= H(T^{-1}TX|T) \\
&\quad \text{como } T \text{ é conhecido, aplicá-lo novamente ou seu inverso não altera a entropia} \\
&= H(X|T) = H(X) \qquad (57) \\
&\quad \text{onde utilizamos que } T \perp\!\!\!\perp X
\end{aligned}
$$

Permutação

O que ocorre se permutarmos as probabilidades?
Seja $p = (p_1, p_2, \ldots, p_n)$, uma distribuição discreta de probabilidade e $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ uma permutação de $1, 2, \ldots, n$.
Considere $p_\sigma = (p_{\sigma_1}, p_{\sigma_2}, \ldots, p_{\sigma_n})$ uma permutação da distribuição p.
Quem será maior? $H(p)$ ou $H(p_\sigma)$?

$$H(p) = -\sum_i p_i \log p_i = -\sum_j p_{\sigma_j} \log p_{\sigma_j} = H(p_\sigma). \tag{58}$$

Sumário

Definição de Entropia

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{59}$$

Entropia Conjunta

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \tag{60}$$

Entropia Condicional

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) \tag{61}$$

Regra da Cadeia

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{62}$$

Limites da Entropia

$$0 \leq H(X) \leq \log n, \text{onde } n \text{ é o tamanho do alfabeto de } X. \tag{63}$$

Entropia do Jogo de Adivinhação

Qual é a melhor estratégia para adivinhar o valor de uma variável aleatória com perguntas sim/não do tipo "$X \in S$?", para algum conjunto $S \subseteq D_X$ (domínio da v.a. X).

Exemplo (Bilmes, 2013)

Seja $X \in D_X = \{x_1, x_2, x_3, x_4, x_5\}$ com probabilidades

x	x_1	x_2	x_3	x_4	x_5
$p(x)$	0.3	0.2	0.2	0.15	0.15

Considere a seguinte estratégia: 1) $X = x_5$? 2) $X = x_4$? 3) $X = x_3$? 4) $X = x_2$? 5) $X = x_1$?

Desta forma faremos 5 perguntas 30% das vezes, 4 perguntas 20% das vezes, etc.

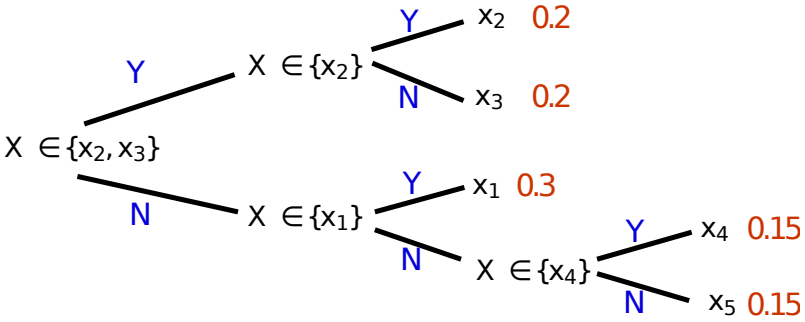
O número médio de perguntas é: $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^\intercal = 3.35$.

Se invertermos a ordem das perguntas teremos: $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^\intercal = 2.65$.

Existe uma estratégia melhor?

Entropia do Jogo de Adivinhação

Considere a estratégia ilustrada abaixo.

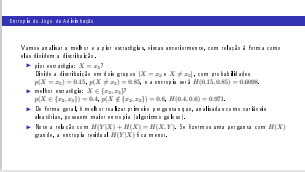


O número médio de perguntas será: $2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3$
Note que $H(X) = 2.271$.
O número médio de perguntas é sempre $\geq H(X)$.

Entropia do Jogo de Adivinhação

Vamos analisar a melhor e a pior estratégias, vistas anteriormente, com relação à forma como elas dividem a distribuição.

- pior estratégia: $X = x_5$?
  Divide a distribuição em dois grupos ($X = x_5$ e $X \neq x_5$), com probabilidades $p(X = x_5) = 0.15$, $p(X \neq x_5) = 0.85$, e a entropia será $H(0.15, 0.85) = 0.6098$.
- melhor estratégia: $X \in \{x_2, x_3\}$?
  $p(X \in \{x_2, x_3\}) = 0.4$, $p(X \notin \{x_2, x_3\}) = 0.6$, $H(0.4, 0.6) = 0.971$.
- De forma geral, é melhor realizar primeiro perguntas que, analisadas como variáveis aleatórias, possuem maior entropia (algoritmo guloso).
- Note a relação com $H(Y|X) + H(X) = H(X, Y)$. Se fizermos uma pergunta com $H(X)$ grande, a entropia residual $H(Y|X)$ fica menor.



Veremos adiante que o algoritmo guloso não é ótimo (entropia mínima).

Intuição sobre Informação Mútua

- Dadas duas variáveis aleatórias X e Y, quanta informação uma possui sobre a outra?
- Conhecendo X, quanto sabemos sobre Y? Conhecendo Y, quanto sabemos sobre X?
- Se as v.a.s são independentes, $X \perp\!\!\!\perp Y$, então conhecer X não nos diz nada sobre Y e vice-versa.
- Como temos uma medida de informação em uma fonte aleatória, $H(X)$, podemos quantificar quanta informação variáveis aleatórias possuem uma sobre as outras. Isto é chamado de informação mútua.

Informação Mútua de Evento

Dado o evento $\{X = x, Y = y\}$, podemos nos perguntar sobre qual é a informação fornecida pelo evento x dado o fato de que o evento y ocorreu. Isto pode ser quantificado da seguinte forma:

$$I(x;y) = \log \frac{p(x|y)}{p(x)} = \underbrace{\log \frac{1}{p(x)}}_{A} - \underbrace{\log \frac{1}{p(x|y)}}_{B} \tag{64}$$

- Primeiro termo A: surpresa de que x ocorreu.
- Segundo termo B: surpresa de que x ocorreu dado que y ocorreu.
- Diferença: diferença entre as duas surpresas, quanto mudou na surpresa de quando não sabíamos y para quando passamos a saber y.

Note que $p(x|x) = 1$, então $I(x;x) = \log 1/p(x) - \log 1 = \log 1/p(x) = I(x)$, então $I(x)$ pode ser visto como uma forma de 'auto-informação'.

Informação Mútua

Informação Mútua é a quantidade média de informação que uma variável aleatória X possui sobre outra v.a. Y e vice-versa.

Definição (Informação Mútua)

$$\begin{aligned} I(X;Y) &= E_{p(x,y)} \log \frac{p(x|y)}{p(x)} = E_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)} \\ &= E_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{65}$$

Informação Mútua e Entropia

Proposição

$$I(X;Y) = H(X) - H(X|Y) \tag{66}$$

Demonstração.

$$\begin{aligned} I(X;Y) &= E\log\frac{p(x|y)}{p(x)} \\ &= E\log\frac{1}{p(x)} - E\log\frac{1}{p(x|y)} \\ &= H(X) - H(X|Y) \end{aligned} \tag{67}$$

□

- Por simetria, temos que $I(X;Y) = H(Y) - H(Y|X)$.
- Como $H(X) \geq 0$ e $H(X|Y) \geq 0$, teremos $I(X;Y) \leq \min(H(X), H(Y))$.

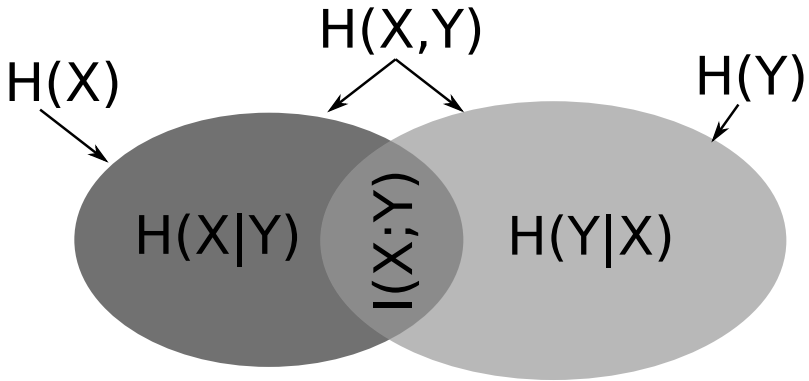
Informação Mútua e Entropia

- Regra da Cadeia da Entropia: $H(X,Y) = H(X) + H(Y|X)$.
- Informação Mútua: $I(X;Y) = H(X) - H(X|Y)$.
- Teremos então:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{68}$$

No próximo slide representamos estas grandezas através de um diagrama. As áreas utilizadas não representam conjuntos no sentido comum, mas representam 'grau de informação' e as intersecções correspondem a sobreposição de informação. Isto é, a interseção consiste em informação fornecida por X e Y.

Informação Mútua e Entropia - Diagrama



Divergência de Kullbach-Leibler

A divergência de Kullbach-Leibler é uma relação fundamental entre duas distribuições probabilísticas sobre um mesmo alfabeto, $p = (p_1, \ldots, p_n)$ e $q = (q_1, \ldots, q_n)$. Esta divergência possui relação importante com a entropia e a informação mútua.

Como podemos medir a 'distância' entre duas distribuições p e q de forma útil? Poderíamos utilizar $D(p, q) = \sum_{i=1}^{n} (p_i - q_i)^2$, mas gostaríamos de ter uma medida de 'distância de informação', isto é, uma distância que nos dê o custo incorrido pelo erro de considerar que uma distribuição é q sendo que na realidade ela é p. Veremos que isto está ligado à insuficiência na compressão. A Divergência de Kullbach-Leibler, definida a seguir, satisfaz estas ideias.

Distância I

Definição (distância)

Seja S um conjunto. Uma função $d : S \times S \to \mathbb{R}$ é chamada **distância** em S se, para todo $x, y \in S$, tivermos:

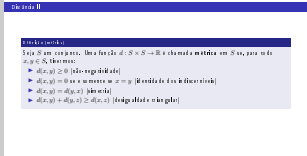
- $d(x, y) \geq 0$ (não-negatividade)
- $d(x, y) = d(y, x)$ (simetria)
- $d(x, x) = 0$ (reflexividade)

Distância II

Definição (métrica)

Seja S um conjunto. Uma função $d : S \times S \to \mathbb{R}$ é chamada **métrica** em S se, para todo $x, y \in S$, tivermos:

- $d(x, y) \geq 0$ (não-negatividade)
- $d(x, y) = 0$ se e somente se $x = y$ (identidade dos indiscerníveis)
- $d(x, y) = d(y, x)$ (simetria)
- $d(x, y) + d(y, z) \geq d(x, z)$ (desigualdade triangular)



Teremos uma semi-métrica se substituirmos a identidade dos indiscerníveis pela reflexividade.

Diferentes formas de medir 'distância' entre distribuições

divergência de Kullback-Leibler: $D_{\mathrm{KL}}(p \parallel q) = \sum p(x) \ln\left(\frac{p(x)}{q(x)}\right)$;

distância de Hellinger: $H^2(p, q) = 2 \sum \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2$;

divergência de Jeffreys: $D_J(p \parallel q) = \sum (p(x) - q(x))\big(\ln p(x) - \ln q(x)\big)$;

divergência α de Chernoff: $D^{(\alpha)}(p \parallel q) = \frac{4}{1-\alpha^2}\left(1 - \sum p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\right)$;

divergência exponencial: $D_e(p \parallel q) = \sum p(x)\big(\ln p(x) - \ln q(x)\big)^2$;

divergência de Kagan: $D_{\chi^2}(p \parallel q) = \frac{1}{2} \sum \frac{(p(x)-q(x))^2}{p(x)}$;

divergência K: $D_{\mathrm{K}}(p \parallel q) = \sum (p(x) - q(x)) \log(p(x)/q(x))$;

divergência de Jensen-Shannon: $D_{\mathrm{JS}}(p \parallel q) = \frac{1}{2} D_{\mathrm{KL}}(p \parallel m) + \frac{1}{2} D_{\mathrm{KL}}(q \parallel m)$, onde $m = \frac{1}{2}(p + q)$.

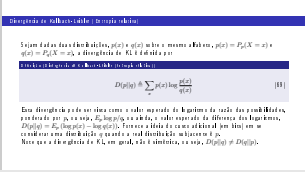
Divergência de Kullbach-Leibler (Entropia relativa)

Sejam dadas duas distribuições, $p(x)$ e $q(x)$ sobre o mesmo alfabeto, $p(x) = P_p(X = x)$ e $q(x) = P_q(X = x)$, a divergência de KL é definida por

Definição (Divergência de Kullbach-Leibler (entropia relativa))

$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{69}$$

Esta divergência pode ser vista como o valor esperado do logaritmo da razão das possibilidades, ponderado por p, ou seja, $E_p \log p/q$, ou ainda, o valor esperado da diferença dos logaritmos, $D(p||q) = E_p (\log p(x) - \log q(x))$. Fornece a ideia do custo adicional (em bits) em se considerar uma distribuição q quando a real distribuição subjacente é p.
Note que a divergência de KL, em geral, não é simétrica, ou seja, $D(p||q) \neq D(q||p)$.


Divergência de Kullbach-Leibler (entropia relativa)

Sejam duas distribuições, $p(x) \in q(x)$ sobre o mesmo alfabeto, $p(x) = P_p(X = x)$ e $q(x) = P_q(X = x)$, a divergência de KL é definida por

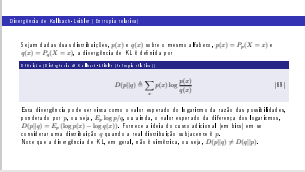
Divergência de Kullbach-Leibler ou Entropia relativa

$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (13)$$

Esta divergência pode ser vista como o valor esperado do logaritmo da razão das probabilidades, ponderado por p, ou seja, $E_p \log p/q$, ou ainda, o valor esperado da diferença dos logaritmos, $D(p||q) = E_p(\log p(x) - \log q(x))$. Tem-se a ideia de como adicional (ou bits) em se considerar uma distribuição q quando a real distribuição subjacente é p.
Nota-se a divergência de KL, em geral, não é simétrica, ou seja, $D(p||q) \neq D(q||p)$.

Utilizando argumentos de limite e continuidade, mostra-se que $0 \log 0 = 0$ e $p \log(p/0) = \infty$. Fazendo estas suposições, teremos $D(p||q) \leq \infty$.
A divergência de KL é uma função dos valores de probabilidade e não dos valores que a variável aleatória assume (assim como a entropia e a informação mútua).

A razão de chances ou razão de possibilidades (em inglês: *odds ratio*) é definida como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo.


Divergência de Kullbach-Leibler (entropia relativa)

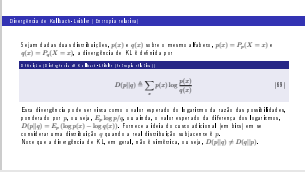
Sejam duas distribuições, $p(x) \in q(x)$ sobre o mesmo alfabeto, $p(x) = P_p(X = x)$ e $q(x) = P_q(X = x)$, a divergência de Kullback-Leibler é definida por

Divergência de Kullbach-Leibler ou Kullback-Leibler ou Kullback-Leibler

$$D(p||q) \triangleq \sum_{x} p(x) \log \frac{p(x)}{q(x)} \quad (13)$$

Esta divergência pode ser vista como o valor esperado do logaritmo da razão das probabilidades, ponderado por p, e o seja, $E_p \log p/q$, e ainda, o valor esperado da diferença dos logaritmos, $D(p||q) = E_p(\log p(x) - \log q(x))$. Tem-se a título de curiosidade (ver bits) em se considerar o seu distribuição q quando a real distribuição subjacente é p.
Note que a divergência de KL, em geral, não é simétrica, ou seja, $D(p||q) \neq D(q||p)$.

(Wikipedia) In statistics, the odds ratio (usually abbreviated "OR") is one of three main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population. If each individual in a population either does or does not have a property "A", (e.g. "high blood pressure"), and also either does or does not have a property "B"(e.g. "moderate alcohol consumption") where both properties are appropriately defined, then a ratio can be formed which quantitatively describes the association between the presence/absence of "A"(high blood pressure) and the presence/absence of "B"(moderate alcohol consumption) for individuals in the population.



This ratio is the odds ratio (OR) and can be computed following these steps:

1. For a given individual that has "B"compute the odds that the same individual has "A"
2. For a given individual that does not have "B"compute the odds that the same individual has "A"
3. Divide the odds from step 1 by the odds from step 2 to obtain the odds ratio (OR). The term "individual"in this usage does not have to refer to a human being, as a statistical population can measure any set of entities, whether living or inanimate.

`http://en.wikipedia.org/wiki/Odds_ratio`

Exemplo

Seja $\mathcal{X} = \{1, 0\}$ e considere duas distribuições p e q em $\mathcal{X}$. Seja $p(0) = 1 - r$, $p(1) = r$, e seja $q(0) = 1 - s$ e $q(1) = s$. Então

$$D(p||q) = (1-r)\log\frac{1-r}{1-s} + r\log\frac{r}{s} \tag{70}$$

e

$$D(q||p) = (1-s)\log\frac{1-s}{1-r} + s\log\frac{s}{r}. \tag{71}$$

Se $r = s$, então $D(p||q) = D(q||p) = 0$. Se $r = \frac{1}{2}$ e $s = \frac{1}{4}$,

$$D(p||q) = \frac{1}{2}\log\frac{1/2}{3/4} + \frac{1}{2}\log\frac{1/2}{1/4} = 1 - \frac{1}{2}\log 3 = 0.2075\text{bits}. \tag{72}$$

$$D(q||p) = \frac{3}{4}\log\frac{3/4}{1/2} + \frac{1}{4}\log\frac{1/4}{1/2} = \frac{3}{4}\log 3 - 1 = 0.1887\text{bits}. \tag{73}$$

Note que, em geral, $D(p||q) \neq D(q||p)$.

Generalização da divergência de KL

A divergência de KL pode ser generalizada para vetores de variáveis aleatórias.
Seja $p(x_1, \ldots, x_N)$ e $q(x_1, \ldots, x_N)$ duas distribuições sobre o vetor $(x_1, x_2, \ldots, x_N)$. A divergência de KL entre p e q é definida por

$$D(p||q) = \sum_{x_1,\ldots,x_N} p(x_1, \ldots, x_N) \log \frac{p(x_1, \ldots, x_N)}{q(x_1, \ldots, x_N)} \tag{74}$$

Divergência de KL e Informação Mútua I

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

$$\begin{aligned} D(\mu_1||\mu_2) &= \sum_{x,y} \mu_1(x,y) \log \frac{\mu_1(x,y)}{\mu_2(x,y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = I(X;Y). \end{aligned} \tag{75}$$

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 v.a.s 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 v.a.s.

Divergência de KL e Informação Mútua II

O produto das distribuições marginais $p(x)p(y)$, onde $p(x) = \sum_y p(x,y)$ e $p(y) = \sum_x p(x,y)$, é uma projeção da distribuição conjunta $p(x,y)$ sobre o conjunto das distribuições independentes. I.e.,

$$p(x)p(y) = \underset{p'(x,y) \backslash p'(x,y)=p'(x)p'(y)}{\operatorname{argmin}} D(p(x,y)||p'(x,y)) \tag{76}$$

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança I

Suponha que tenhamos uma v.a. $\mathbf{X} = (X_1, \ldots, X_N)$ com uma distribuição subjacente p que depende de um parâmetro θ (modelo hipotético). Queremos definir um estimador $\hat{\theta} = T(X_1, \ldots, X_N)$ para este parâmetro θ, dadas as observações $x_1, \ldots, x_N$. Um bom estimador para o parâmetro desconhecido θ é aquela que maximiza a verossimilhança $L(\theta)$ do parâmetro, dada a observação dos dados,

$$L(\theta) = \Pr(X_1 = x_1, \ldots, X_N = x_N) = p(x_1|\theta) \ldots p(x_N|\theta) = \prod_{n=1}^{N} p(x_n|\theta). \tag{77}$$

Como a função logaritmo é monotônica crescente, maximizar $L(\theta)$ é equivalente a maximiza $l(\theta) = \log L(\theta)$,

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta). \tag{78}$$

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança II

A estimativa de máxima verossimilhança (MLE, *maximum likelihood estimator*) de θ é dada por

$$\hat{\theta}_{\text{mle}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta\,;\, x_1, \ldots, x_N). \tag{79}$$

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança III

Chamaremos de $\hat{p}$ a distribuição empírica. Seja $x_1, \ldots, x_N \in \mathcal{X}$, N observações i.i.d. de uma variável aleatória X. A distribuição empírica será dada por

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n), \tag{80}$$

onde δ é a função de Dirac.

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 $D_{\mathrm{KL}}(\hat{p} \parallel p_\theta)$.

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança IV

$$
\begin{aligned}
D_{\mathrm{KL}}(\hat{p} \parallel p_\theta) &= \sum_{x \in \mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} \\
&= -H(\hat{p}) - \sum_{x \in \mathcal{X}} \hat{p}(x) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{N} \sum_{x \in \mathcal{X}} \sum_{n=1}^{N} \delta(x - x_n) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n). \qquad (81)
\end{aligned}
$$

O segundo termo é o oposto do logaritmo da verossimilhança de $p_\theta(x)$.

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança V

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

$$\begin{aligned}\hat{\theta}_n &= \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_{n=1}^{N} p_\theta(x_n) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_\theta(x_n) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} -\log p_\theta(x_n). \end{aligned} \tag{82}$$

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).

Informação Mútua Condicionada a um evento

A informação pode se alterar se for condicionada a um evento de uma terceira variável aleatória $\{Z = z\}$, e isto é denotado por $I(X;Y|Z = z)$, onde X, Y, Z são variáveis aleatórias. Dada a distribuição $p(x, y, z)$, a informação mútua condicionada ao evento específico $\{Z = z\}$ é dada por

$$I(X;Y|Z = z) = \sum_{x,y} p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}. \tag{83}$$

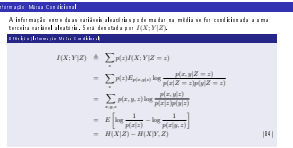
Obs. Fazemos as seguintes alterações sobre a informação mútua padrão: $p(x, y) \to p(x, y|z)$, $p(x) \to p(x|z)$ e $p(y) \to p(y|z)$.

Informação Mútua Condicional

A informação entre duas variáveis aleatórias pode mudar na média se for condicionada a uma terceira variável aleatória. Será denotada por $I(X;Y|Z)$.

Definição (Informação Mútua Condicional)

$$\begin{aligned} I(X;Y|Z) &\triangleq \sum_z p(z) I(X;Y|Z=z) \\ &= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x,y|Z=z)}{p(x|Z=z)p(y|Z=z)} \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= E\left[\log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y,z)}\right] \\ &= H(X|Z) - H(X|Y,Z) \end{aligned} \tag{84}$$



$$I(X;Y) = H(X) - H(X|Y) \tag{85}$$

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{86}$$

Regra da Cadeia para Informação Mútua

Proposição

$$I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y | X_1, X_2, \ldots, X_{i-1}) \tag{87}$$

Exemplo: $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1)$

Demonstração.

$$\begin{aligned} I(X_1, \ldots, X_N; Y) &= H(X_1, \ldots, X_N) - H(X_1, \ldots, X_N | Y) \\ &= \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}) - \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}, Y) \\ &= \sum_{i=1}^{N} I(X_i; Y | X_1, \ldots, X_{i-1}) \end{aligned} \tag{88}$$

□

Entropia Relativa Condicional - divergência de KL

Definição

Para pmf conjuntas $p(x, y)$ e $q(x, y)$, a entropia relativa condicional é definida como

$$\begin{aligned} D(p(y|x)||q(y|x)) &\triangleq \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x,y} p(x, y) \log \frac{p(y|x)}{q(y|x)} \\ &= \mathrm{E}_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}, \end{aligned} \tag{89}$$

é o valor esperado das entropias relativas entre as pmfs condicionais $p(y|x)$ e $q(y|x)$, tomando o valor esperado sobre a distribuição de massa $p(x)$.

Regra da Cadeia para divergência de KL

Proposição

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \tag{90}$$

Demonstração.

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(x,y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} + \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} \end{aligned} \tag{91}$$

□

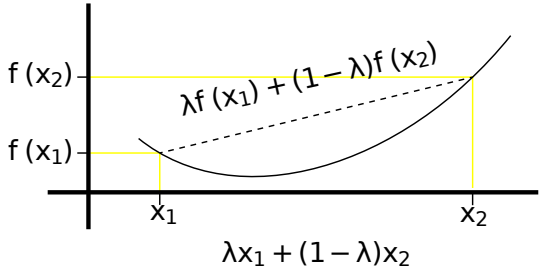
Funções Convexas

Definição

Dizemos que f é convexa em (a,b) se para todo $x_1, x_2 \in (a,b)$, $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{92}$$

Exemplos: 1) $f(x) = x^2$ 2) $f(x) = e^x$ 3) $x \log x$, $x \geq 0$.



- f é estritamente convexa se a igualdade for verdadeira apenas para $\lambda = 0$ ou $\lambda = 1$.

Derivada Segunda e Convexidade I

Teorema (derivada segunda e 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

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \tag{93}$$

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

...

Derivada Segunda e Convexidade II

Demonstração.

continuação...

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

$$\begin{aligned} f(x_1) &\geq f(x_0) + f'(x_0)(x_1 - \lambda x_1 - (1-\lambda)x_2) \\ &= f(x_0) + f'(x_0)((1-\lambda)(x_1 - x_2)) \end{aligned} \tag{94}$$

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

$$\begin{aligned} f(x_2) &\geq f(x_0) + f'(x_0)(x_2 - \lambda x_1 - (1-\lambda)x_2) \\ &= f(x_0) + f'(x_0)(\lambda(x_2 - x_1)) \end{aligned} \tag{95}$$

...

Derivada Segunda e Convexidade III

Demonstração.

continuação...

Somando λ 94 com $(1-\lambda)$ 95, teremos

$$\begin{aligned} \lambda f(x_1) + (1-\lambda) f(x_2) &\geq \lambda f(x_0) + \lambda f'(x_0)((1-\lambda)(x_1 - x_2)) + \\ &\quad (1-\lambda) f(x_0) + (1-\lambda) f'(x_0)(\lambda(x_2 - x_1)) \\ &\geq f(x_0) = f(\lambda x_1 + (1-\lambda) x_2) \end{aligned} \tag{96}$$

□

Desigualdade de Jensen

Teorema (Jensen)

Seja f uma função convexa e X uma variável aleatória, então

$$Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f\left(\sum_x xp(x)\right) \tag{97}$$

Se f é estritamente convexa, então $\{Ef(X) = f(E(X))\} \Rightarrow \{X = EX\}$, o que significa que X é uma v.a. constante.

Desigualdade de Jensen - demonstração I

- Para uma distribuição de massa com apenas dois pontos.

$$E[f(X)] = p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) = f(EX) \tag{98}$$

  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.

Demonstração.

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 $i = 1, 2, \ldots, k-1$. ...

Desigualdade de Jensen - demonstração II

Demonstração.

continuação...

Desta forma, teremos

$$\begin{aligned} E[f(X)] &= \sum_{i=1}^{k} p_i f(x_i) \\ &= \sum_{i=1}^{k-1} (1-p_k) p_i' f(x_i) + p_k f(x_k) \\ &= (1-p_k) \sum_{i=1}^{k-1} p_i' f(x_i) + p_k f(x_k) \end{aligned}$$

...

Desigualdade de Jensen - demonstração III

Demonstração.

continuação...

Poderemos utilizar a hipótese de indução, já que

$$\sum_{i=1}^{k-1} p'_i = \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} = \frac{1-p_k}{1-p_k} = 1. \tag{99}$$

Então

$$\begin{aligned} E[f(X)] &= \dots \\ &\geq (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) + p_k f(x_k) \end{aligned}$$

...

Desigualdade de Jensen - demonstração IV

Demonstração.

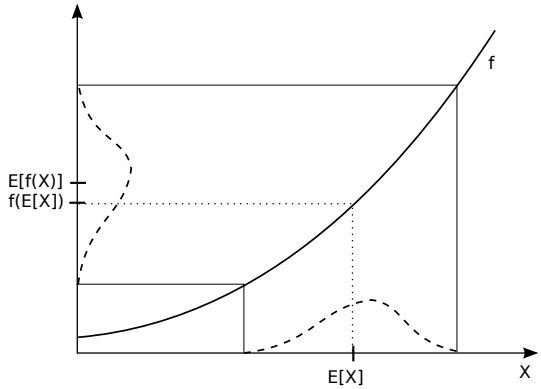
continuação...

Pela definição de convexidade, teremos

$$\begin{aligned} E[f(X)] &= \dots \\ &\geq f\left((1-p_k)\sum_{i=1}^{k-1} p_i' x_i + p_k x_k\right) \\ &= f\left(\sum_{i=1}^{k} p_i x_i\right) \end{aligned} \tag{100}$$

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. □

Desigualdade de Jensen - demonstração gráfica



O mapeamento feito pela função convexa f aumenta gradativamente o estiramento da distribuição mapeada por f com o aumento dos valores de X. Desta forma, o valor esperado da distribuição mapeada por f tende a possuir um valor maior que o mapeamento por f do valor esperado da distribuição.

A divergência de KL é não-negativa

Lema

$$D(p||q) \geq 0 \text{ com igualdade se e somente se } p(x) = q(x) \forall x. \tag{101}$$

A divergência de KL é não-negativa

Demonstração.

Mostre que $-D(p||q) \leq 0$. Seja $S_p = \{x : p(x) > 0\} = \sup(p)$, então

$$
\begin{aligned}
-D(p||q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in S_p} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in S_p} p(x) \log \frac{q(x)}{p(x)} = E \log \frac{q(X)}{p(X)} \\
&\quad \text{utilizando a desigualdade de Jensen} \\
&\leq \log \left(E \frac{q(X)}{p(X)} \right) = \log \left(\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} \right) \\
&= \log \left(\sum_{x \in S_p} q(x) \right) \leq \log \left(\sum_x q(x) \right) = \log 1 = 0 \qquad (102)
\end{aligned}
$$

□

A divergência de KL é não-negativa

- Note que $\log x$ é estritamente côncavo.
- Então, a igualdade $\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} = \log \left(\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} \right)$ significa $Z = EZ$ com $Z = q(X)/p(X)$, então Z é uma variável aleatória constante.
- A única constante válida, com p e q sendo distribuições de probabilidade é $Z = 1$ ou $p(x) = q(x)$.
- Então, se $p(x) = q(x)$ teremos $D(p||q) = 0$ e vice-versa.

A informação mútua é não-negativa

Proposição

$$I(X;Y) \geq 0 \text{ e } I(X;Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y. \tag{103}$$

Demonstração.

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0 \tag{104}$$

□

teremos igualdade se $p(x,y) = p(x)p(y)$, o que é também condição para independência.

Informação Mútua

- $I(X;Y)$ mede o 'grau de dependência' entre X e Y.
- Temos $0 \leq I(X;Y) \leq \min(H(X), H(Y))$.
- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.
- Se $X \perp\!\!\!\perp Y$, então $I(X;Y) = 0$, já que em tal caso $H(X|Y) = H(X)$ e $H(Y|X) = H(Y)$.
- Se $X = Y$, então $I(X;Y) = H(X) = H(Y)$ já que em tal caso $H(Y|X) = H(X|Y) = 0$.

Limite Superior para a Entropia I

Teorema

$H(X) \leq \log |\mathcal{X}|$, onde $|\mathcal{X}|$ denota o número de elementos da extensão de X (a cardinalidade do domínio), com igualdade se e somente se X possuir distribuição uniforme.

Limite Superior para a Entropia II

Demonstração.

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

$$\begin{aligned} D(p||u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= \sum_{x\in\mathcal{X}} p(x)\log p(x) + \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} \\ &= -H(X) + \log|\mathcal{X}|\sum_{x\in\mathcal{X}} p(x) \\ &= \log|\mathcal{X}| - H(X) \end{aligned} \tag{105}$$

...

Limite Superior para a Entropia III

Demonstração.

continuação...

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

$$D(p||u) = \log |\mathcal{X}| - H(X) \geq 0 \tag{106}$$

e assim

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

□

Condicionar Reduz Entropia

Comparando $H(X)$ com $H(X|Y)$, conhecendo Y, na média, pode nos dizer algo sobre X reduzindo a entropia.

Proposição

$$H(X|Y) \leq H(X) \text{ e } H(X|Y) = H(X) \text{ se e somente se } X \perp\!\!\!\perp Y. \tag{108}$$

Demonstração.

$$0 \leq I(X;Y) = H(X) - H(X|Y) \tag{109}$$

□

Poderíamos ter $H(X|Y=y) > H(X)$, mas, na média, $\sum_y p(y)H(X|Y=y) \leq H(X)$.

Limite da Entropia para um Conjunto de V.A.

A entropia de um conjunto de variáveis aleatórias é maior quando as variáveis aleatórias são independentes, há menor redundância entre elas.

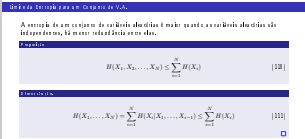
Proposição

$$H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \tag{110}$$

Demonstração.

$$H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \tag{111}$$

□


Limite da Entropia para um Conjunto de V.A.

A entropia de um conjunto de variáveis aleatórias é maior quando as variáveis aleatórias são independentes, há menor redundância entre elas.

Propriedade

$$H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \qquad (112)$$

Demonstração

$$H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \qquad (111)$$

$$\sum_{i=1}^{N} H(X_i|X_{\neg i}) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N) \leq H(X_1, \ldots, X_N) \qquad (112)$$

Limites da Independência na Entropia

A proposição 6 para duas variáveis é da forma

$$H(X_1, X_2) \leq H(X_1) + H(X_2) \tag{113}$$

Note que a igualdade na Equação 110 é alcançada quando todas as variáveis são mutuamente independentes, isto é, quando $X_i \perp\!\!\!\perp X_j \forall i, j$.

Condicionamento e Informação Mútua

- Se $X \perp\!\!\!\perp Y|Z$ então $I(X;Y|Z) = 0$. Por exemplo, $X \perp\!\!\!\perp Y|Z$ quando $X \to Z \to Y$.
- Alternativamente, se $Z = Y$, então $I(X;Y|Z) = 0$.
- Podemos ter $I(X;Y) > I(X;Y|Z)$.
- Por outro lado, se $Z = X + Y$ e $X \perp\!\!\!\perp Y$, então $I(X;Y) = 0$ mas $I(X;Y|Z) > 0$.
- Não existe uma relação genérica entre informação mútua e informação mútua condicional.

Relações de H

$$H(X) = EI(X) = -\sum_{x} p(x) \log p(x) \tag{114}$$

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \tag{115}$$

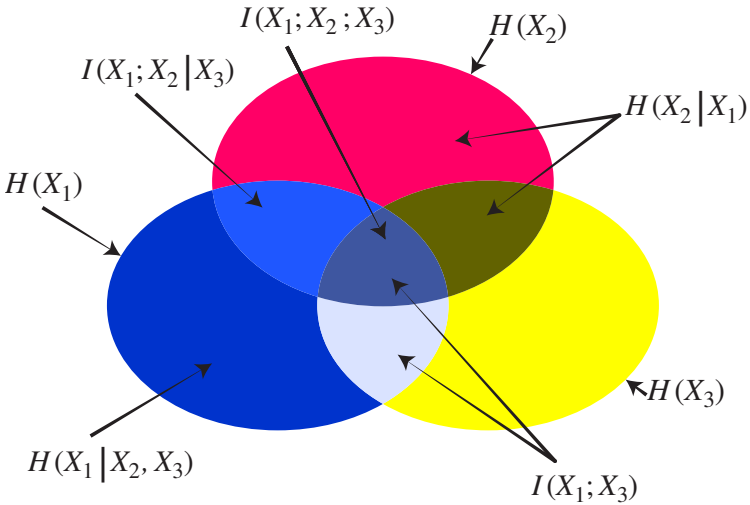
$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) \tag{116}$$

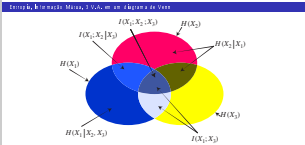
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{117}$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{118}$$

$$0 \leq H(X) \leq \log n, \text{ onde } n \text{ é o tamanho do alfabeto de } X. \tag{119}$$

Entropia, Informação Mútua, 3 V.A. em um diagrama de Venn





- $I(X_1;X_2) = I(X_1;X_2|X_3) + I(X_1;X_2;X_3)$.
- $I(X_1;X_2) \gtreqless I(X_1;X_2|X_3)$, mas isto nunca será negativo.
- Então, $I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3)$ pode ser negativo.
- $I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3) = I(X_2;X_3) - I(X_2;X_2|X_1) = I(X_3;X_1) - I(X_3;X_1|X_2)$
- $I(X_1;X_2;X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1;X_2) - H(X_2;X_3) - H(X_3;X_1) + H(X_1,X_2,X_3)$

Revisão I

- divergência de KL: $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$
- informação mútua: $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$
- informação mútua condicional:
  $I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = H(X|Z) - H(X|Y,Z)$
- regra da cadeia da informação mútua:
  $I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1})$
- entropia relativa condicional: $D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)}$
- regra da cadeia da divergência de KL:
  $D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$
- Jensen: f convexa $\Rightarrow Ef(X) = \sum_x p(x) f(x) \geq f(EX) = f\left(\sum_x px(x)\right)$
- não negatividade da divergência de KL: $D(p||q) \geq 0$, $D(p||q) = 0 \Leftrightarrow p = q$.
- não negatividade da informação mútua: $I(X;Y) \geq 0$, $I(X;Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

Revisão II

- condicionar reduz a entropia: $H(X) \geq H(X|Y)$, $H(X) = H(X|Y) \Leftrightarrow X \perp\!\!\!\perp Y$.
- limite da independência em H: $H(X_1, \ldots, X_N) \leq \sum_i H(X_i)$, com igualdade sse todos X_i forem independentes

Medida de Informação

Veremos as correspondências entre a medida de informação de Shannon (e suas manipulações) com a teoria de conjuntos. A utilização dos diagramas de informação podem ser utilizadas para simplificar várias demonstrações em teoria da informação.

Medida de Informação

- Temos um conjunto de variáveis aleatórias: $X_1, X_2, \ldots, X_n$.
- Para cada variável aleatória associamos um conjunto $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

Definição (campo (field))

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 (átomo)

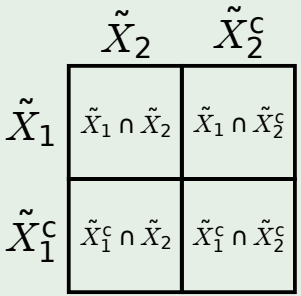
Os átomos de $\mathcal{F}_n$ são os conjuntos da forma $\cap_{i=1}^{n} Y_i$, onde Y_i é $\tilde{X}_i$ ou $\tilde{X}_i^c$, o complemento de $\tilde{X}_i$.

Átomos

átomos - $n = 2$

Para $n = 2$, teremos os conjuntos $\tilde{X}_1, \tilde{X}_2$ e seus complementos, respectivamente, $\tilde{X}_1^c, \tilde{X}_2^c$. Existirão 4 átomos:

1) $\tilde{X}_1 \cap \tilde{X}_2$,
2) $\tilde{X}_1 \cap \tilde{X}_2^c$,
3) $\tilde{X}_1^c \cap \tilde{X}_2$, e
4) $\tilde{X}_1^c \cap \tilde{X}_2^c$

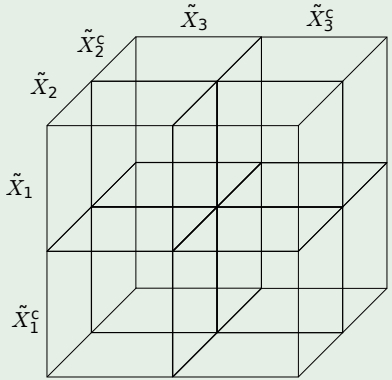


Átomos

átomos - $n = 3$

Para $n = 3$, 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:

1) $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3$,
2) $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c$,
3) $\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3$,
4) $\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3^c$,
5) $\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3$,
6) $\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3^c$,
7) $\tilde{X}_1^c \cap \tilde{X}_2^c \cap \tilde{X}_3$, e
8) $\tilde{X}_1^c \cap \tilde{X}_2^c \cap \tilde{X}_3^c$



Campo e Átomos

- existem 2^n átomos;
- existem 2^{2^n} conjuntos no campo $\mathcal{F}_n$;
- todos os átomos em $\mathcal{F}_n$ são disjuntos;
- todo conjunto em $\mathcal{F}_n$ pode ser expresso de forma única como uma união de um subconjunto dos átomos em $\mathcal{F}_n$.

Medida com sinal

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 (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 $\mathcal{F}_n$,

$$\mu(A \cup B) = \mu(A) + \mu(B). \tag{120}$$

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

Medida com sinal

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.

Medida com sinal

$n = 2$

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

$$\begin{aligned} \mu(\tilde{X}_1) &= \mu((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \\ &= \mu(\tilde{X}_1 \cap \tilde{X}_2) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c). \end{aligned} \tag{121}$$

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

$$\mu(\tilde{X}_1 \setminus \tilde{X}_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c). \tag{122}$$

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

$$\begin{aligned} \mu(\tilde{X}_1 \cup \tilde{X}_2) &= \mu((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c) \cup (\tilde{X}_1^c \cap \tilde{X}_2)) \\ &= \mu(\tilde{X}_1 \cap \tilde{X}_2) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1^c \cap \tilde{X}_2) \end{aligned} \tag{123}$$

Correspondência com a informação de Shannon I

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 4.

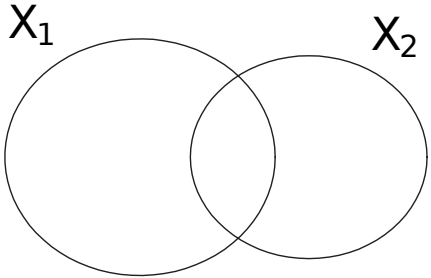


Figura 4: Diagrama de informação para X_1 e X_2.

Correspondência com a informação de Shannon II

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,

$$\tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \Omega^c = \emptyset. \tag{124}$$

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

$$H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2), I(X_1; X_2). \tag{125}$$

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

$$\mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2), \tag{126}$$
$$\mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1), \tag{127}$$
$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2). \tag{128}$$

Correspondência com a informação de Shannon III

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

$$\mu^*(\tilde{X}_1 \cup \tilde{X}_2) = H(X_1, X_2), \quad (129)$$
$$\mu^*(\tilde{X}_1) = H(X_1), \quad (130)$$
$$\mu^*(\tilde{X}_2) = H(X_2). \quad (131)$$

Por exemplo, a Equação 129 pode ser verificada

$$\begin{aligned} \mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= \mu^*((\tilde{X}_1 \setminus \tilde{X}_2) \cup (\tilde{X}_2 \setminus \tilde{X}_1) \cup (\tilde{X}_1 \cap \tilde{X}_2)) \\ &= \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) + \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) + \mu^*(\tilde{X}_1 \cap \tilde{X}_2) \\ &= H(X_1|X_2) + H(X_2|X_1) + I(X_1; X_2) \\ &= H(X_1, X_2). \end{aligned} \quad (132)$$

Correspondência com a informação de Shannon IV

A Equação 130 também pode ser facilmente verificada

$$\begin{aligned}\mu^*(\tilde{X}_1) &= \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \\ &= \mu^*(\tilde{X}_1 \cap \tilde{X}_2) + \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) \\ &= I(X_1; X_2) + H(X_1|X_2) = H(X_1). \end{aligned} \tag{133}$$

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

$$H/I \leftrightarrow \mu^* \tag{134}$$
$$, \leftrightarrow \cup \tag{135}$$
$$; \leftrightarrow \cap \tag{136}$$
$$| \leftrightarrow \setminus \tag{137}$$

- obs.: 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.

Desigualdade da soma de logaritmos

Teorema (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*

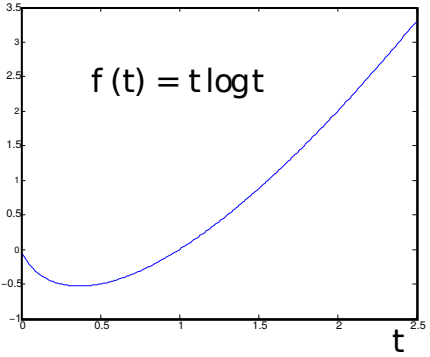
$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{138}$$

e teremos igualdade sse $a_i/b_i = c = const.$.

- Relembrando: $0 \log 0 = 0$, $a \log a/0 = \infty$ para $a > 0$, e $0 \log 0/0 = 0$.
- A desigualdade da soma de logaritmos é utilizada para demonstrar algumas propriedades importantes.

Desigualdade da soma de logaritmos

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



Desigualdade da soma de logaritmos I

Demonstração.

- Dada f convexa, a desigualdade de Jensen diz que

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \text{ com } \alpha_i \geq 0 \text{ e } \sum_i \alpha_i = 1 \tag{139}$$

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

...

Desigualdade da soma de logaritmos II

Demonstração.

continuação...

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

$$\begin{aligned} \sum_i \alpha_i f(t_i) &\geq f\left(\sum_i \alpha_i t_i\right) \\ \sum_i \left(\frac{b_i}{\sum_j b_j} f\left(\frac{a_i}{b_i}\right)\right) &\geq f\left(\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\right) \end{aligned}$$

...

Desigualdade da soma de logaritmos III

Demonstração.

continuação...

$$\begin{aligned}
\sum_i \left(\frac{b_i}{\sum_j b_j} f\left(\frac{a_i}{b_i}\right)\right) &\geq f\left(\sum_i \frac{b_i}{\sum_j b_j}\frac{a_i}{b_i}\right) \\
\frac{1}{\sum_j b_j}\sum_i \left(b_i \frac{a_i}{b_i}\log\frac{a_i}{b_i}\right) &\geq \left(\sum_i \frac{a_i}{\sum_j b_j}\right)\log\sum_i \frac{a_i}{\sum_j b_j} \\
\sum_i a_i \log\frac{a_i}{b_i} &\geq \left(\sum_i a_i\right)\log\sum_i \frac{a_i}{\sum_j b_j} \\
\sum_i a_i \log\frac{a_i}{b_i} &\geq \sum_i a_i \log\frac{\sum_i a_i}{\sum_j b_j} \qquad (140)
\end{aligned}$$

□

Divergência é não negativa

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

Demonstração.

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &\geq \left(\sum_x p(x) \right) \log \frac{\sum_x p(x)}{\sum_x q(x)} \\ &= 1 \log \frac{1}{1} = 0 \end{aligned} \tag{141}$$

□

A Entropia Relativa é Convexa no Par I

Teorema

Seja (p_1, q_1) *e* (p_2, q_2) *dois pares de função massa probabilidade, então*

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2), \tag{142}$$

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

A Entropia Relativa é Convexa no Par II

Demonstração.

Pela definição da divergência de KL, temos

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) = \sum_x (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \quad (143)$$

cada termo do somatório é da forma

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} = \left(\sum_i a_i \right) \log \frac{\sum_i a_i}{\sum_i b_i} \quad (144)$$

...

A Entropia Relativa é Convexa no Par III

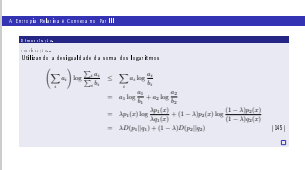
Demonstração.

continuação...

Utilizando a desigualdade da soma dos logaritmos

$$\begin{aligned}\left(\sum_i a_i\right)\log\frac{\sum_i a_i}{\sum_i b_i} &\leq \sum_i a_i \log\frac{a_i}{b_i} \\ &= a_1\log\frac{a_1}{b_1} + a_2\log\frac{a_2}{b_2} \\ &= \lambda p_1(x)\log\frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x)\log\frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)} \\ &= \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2) \qquad (145)\end{aligned}$$

□



- Note que podemos fazer $q_1 = q_2$ e desta forma obteremos convexidade apenas em p.
- Este é o fundamento para o procedimento de minimização alternada, que é um caso especial do algoritmo de EM (maximização da esperança), para o cálculo da função de taxa de distorção, e para o cálculo da função geral de capacidade de canal.

A Entropia é Concava I

Teorema

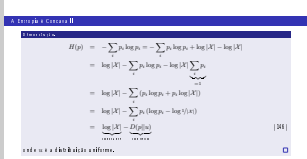
$H(p)$ é uma função concava de p.

A Entropia é Concava II

Demonstração.

$$\begin{aligned} H(p) &= -\sum_i p_i \log p_i = -\sum_i p_i \log p_i + \log|\mathcal{X}| - \log|\mathcal{X}| \\ &= \log|\mathcal{X}| - \sum_i p_i \log p_i - \log|\mathcal{X}| \underbrace{\sum_i p_i}_{=1} \\ &= \log|\mathcal{X}| - \sum_i (p_i \log p_i + p_i \log|\mathcal{X}|) \\ &= \log|\mathcal{X}| - \sum_i p_i \left(\log p_i - \log {}^1\!/_{|\mathcal{X}|}\right) \\ &= \underbrace{\log|\mathcal{X}|}_{\text{constante}} - \underbrace{D(p||u)}_{\text{convexo}} \end{aligned} \tag{146}$$

onde u é a distribuição uniforme. □



Podemos ver a entropia como a similaridade com a distribuição uniforme. Quanto maior a entropia, mais próximo estaremos da distribuição uniforme.

Consequências para a Informação Mútua I

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.

Consequências para a Informação Mútua II

Demonstração.

- $I(X;Y)$ é uma função côncava de $p(x)$ para $p(y|x)$ fixo

$$\begin{aligned} I(X;Y) &= D(p(x,y)||p(x)p(y)) \text{ (definição)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \text{ (definição)} \\ &= \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \end{aligned} \tag{147}$$

Se $p(y|x)$ é constante, então a informação mútua é função de $p(x)$

$$I_{p(x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \tag{148}$$

...

Consequências para a Informação Mútua III

Demonstração.

continuação...

$$I_{p(x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \tag{149}$$

Utilizando a propriedade da convexidade da divergência de Kullback-Leibler

$$I_{\lambda p_1(x)+(1-\lambda)p_2(x)}(X;Y) \geq \lambda I_{p_1(x)}(X;Y) + (1-\lambda) I_{p_2(x)}(X;Y) \tag{150}$$

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

...

Consequências para a Informação Mútua IV

Demonstração.

continuação...

- $I(X;Y)$ é uma função convexa de $p(y|x)$ para $p(x)$ fixo

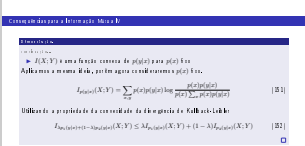
Aplicamos a mesma ideia, porém agora consideraremos $p(x)$ fixo.

$$I_{p(y|x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \tag{151}$$

Utilizando a propriedade da convexidade da divergência de Kullback-Leibler

$$I_{\lambda p_1(y|x)+(1-\lambda)p_2(y|x)}(X;Y) \leq \lambda I_{p_1(y|x)}(X;Y) + (1-\lambda) I_{p_2(y|x)}(X;Y) \tag{152}$$

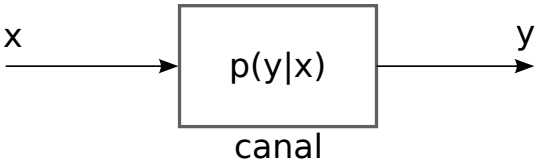
□



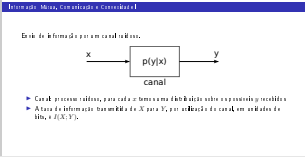
Estes resultados serão importantes para a capacidade de canal e vários outras otimizações envolvendo informação mútua e distribuições.

Informação Mútua, Comunicação e Convexidade I

Envio de informação por um canal ruidoso.



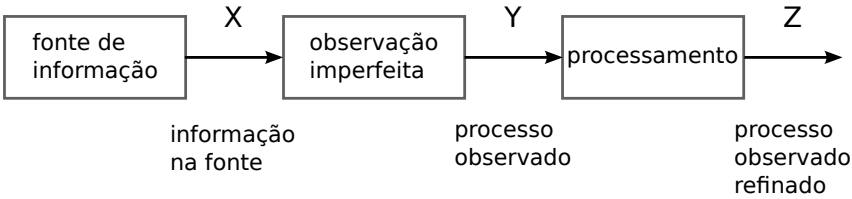
- Canal: processo ruidoso, para cada x temos uma distribuição sobre os possíveis y recebidos
- A taxa de informação transmitida de X para Y, por utilização do canal, em unidades de bits, é $I(X;Y)$.



- Embaralhando $p(x)$ não pode diminuir (pode aumentar ou não alterar) a transmissão de informação para um canal fixo, com relação à original mistura de taxas (ver Equação 150).
- Embaralhando $p(y|x)$ para um canal ruidoso e uma fonte fixa, podemos apenas não aumentar (pode reduzir ou manter constante) a taxa de transmissão, em relação à original mistura de taxas (ver Equação 152).

Desigualdade do Processamento de Dados

Dada uma fonte de informação, é possível utilizar alguma forma de processamento de dados de forma a obter mais informação sobre esta fonte?



Desigualdade do Processamento de Dados

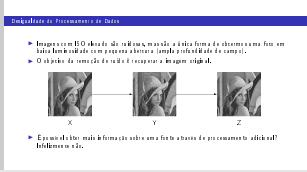
- Imagens com ISO elevado são ruidosas, mas são a única forma de obtermos uma foto em baixa luminosidade com pequena abertura (ampla profundidade de campo).
- O objetivo da remoção de ruído é recuperar a imagem original.

- É possível obter mais informação sobre uma fonte através de processamento adicional? Infelizmente não.

Desigualdade do Processamento de Dados

- Imagens com ISO elevado são ruidosas, mas são a única forma de obtermos uma foto em baixa luminosidade com pequena abertura (ampla profundidade de campo).
- O objetivo da remoção de ruído é recuperar a imagem original.

- É possível obter mais informação sobre uma fonte através de processamento adicional? Infelizmente não.



Profundidade de campo descreve até que ponto objetos que estão mais ou menos perto do plano de foco aparentam estar nítidos.

Regra geral, quanto menor for a abertura do diafragma/íris (maior o valor f/x), para uma mesma distância do objecto fotografado, maior será a distância do plano de foco a que os objetos podem estar enquanto permanecem nítidos.

Cadeia de Markov I

Definição (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

$$p(x, y, z) = p(x)p(y|x)p(z|y) \tag{153}$$

- $X \to Y \to Z$ sse X e Z são condicionalmente independentes dado Y ($X \perp\!\!\!\perp Z|Y$). Isto é,

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = p(x|y)p(z|y) \quad \forall x, y, z \tag{154}$$

Cadeia de Markov II

- $X \to Y \to Z$ implica em $Z \to Y \to X$

Demonstração.

$$\begin{aligned} p(x,y,z) &= p(x)p(y|x)p(z|y) = p(x,y)p(z|y) \\ &= \frac{p(x,y)p(z|y)p(y)}{p(y)} = p(x|y)p(y,z) \\ &= p(x|y)p(y|z)p(z) \end{aligned} \tag{155}$$

□

- Se $Z = f(Y)$, então $X \to Y \to Z$ (i.e., X, Y e Z formam uma cadeia de Markov). $f(\cdot)$ pode ser aleatória ou determinística. X é irrelevante para determinar Z quando Y é dado.

Desigualdade do Processamento de Dados I

Teorema (Desigualdade do Processamento de Dados)

Se $X \to Y \to Z$ *então*

$$I(X;Y) \geq I(X;Z) \tag{156}$$

- Na cadeia de Markov, as setas correspondem ao processamento e as variáveis aleatórias correspondem aos dados.
- O processamento pode ser aleatório ou determinístico.
- A desigualdade de processamento de dados diz que ao efetuar mais processamento dos dados, só é possível perder informação sobre a fonte original, quando medida pela informação mútua.

Desigualdade do Processamento de Dados II

Demonstração.

Utilizando a regra da cadeia da informação mútua, teremos

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned} \tag{157}$$

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

$$I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} = I(X;Y) + \cancelto{0}{I(X;Z|Y)} \tag{158}$$

Então

$$I(X;Z) \leq I(X;Y) \tag{159}$$

□

Desigualdade do Processamento de Dados III

- Teremos igualdade sse $I(X;Y|Z) = 0$ (i.e. $X \to Z \to Y$).
- De forma similar, podemos mostrar que $I(Y;Z) \geq I(X;Z)$.

Corolário

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

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

$$\underbrace{I(X;Z)}_{\geq 0} + I(X;Y|Z) = I(X;Y) + \cancelto{0}{I(X;Z|Y)} \tag{160}$$

então

$$I(X;Y|Z) \leq I(X;Y) \tag{161}$$

Desigualdade do Processamento de Dados IV

- Processamento pode apenas perder informação sobre X. Quando X é a fonte e Y o receptor, nenhum processamento irá aumentar a informação sobre X.
- Considere o reconhecimento de padrões: X é um objeto, Y é uma lista de características e $f(Y)$ é processamento subsequente. Então, qualquer processamento subsequente poderá apenas reduzir a informação sobre o objeto.
- Como funciona então as técnicas de remoção de ruído em imagens ou áudio?
  - As técnicas supõem o conhecimento de algumas informações sobre a imagem original, ou seja, utilizam conhecimento *a priori* para processar a imagem.

Corolário

Se $X \to Y \to Z$, *então* $I(X;Y|Z) \leq I(X;Y)$. *I.e.,*
$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \leq H(X) - H(X|Y)$.

Estatística T I

Exemplo

Seja $X_1, X_2, \ldots, X_N$, $X_i \in \{0, 1\}$ uma sequência i.i.d. de arremessos de moeda, $p(X = 1) = \theta = 1 - P(X = 0)$.
Faça $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$, contagem do número de *caras*.
Dizemos que T é uma **estatística** da amostra.

- De forma geral, uma estatística é uma função de uma coleção de variáveis aleatórias (e.g., uma média empírica, uma variância empírica, ou um máximo empírico, etc)
- Uma estatística é por sua vez uma v.a.
- Uma boa estatística possui informação útil sobre as amostras, enquanto uma estatística ruim não (por exemplo $T(X_1, \ldots, X_N) = X_1$).
- As estatísticas costumam ser chamadas de 'características' no contexto de reconhecimento de padrões e aprendizado de máquina.

Ensaio de Bernoulli I

- Considere a estatística de contagem citada anteriormente.
- Uma vez que sabemos a estatística, a probabilidade de uma sequência pode ser expressa sem fazer referência à θ (parâmetro que caracteriza a distribuição).

$$p(x_1,\ldots,x_N|T(x_1,\ldots,x_N),\theta) = p(x_1,\ldots,x_N|T(x_1,\ldots,x_N)) = \begin{cases} \frac{1}{\binom{N}{k}} & \sum_i x_i = k \\ 0 & \text{caso contrário} \end{cases} \quad (162)$$

- Em outras palavras: $X_{1:N} \perp\!\!\!\perp \theta|T(X_{1:N})$.
- Isto implica na cadeia de Markov: $\theta \to T(X_{1:N}) \to X_{1:N}$
- 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: $\theta \to X_{1:N} \to T(X_{1:N})$

Desigualdade de Processamento de Dados e Estatística I

- cadeia de Markov (A): $\theta \to T(X_{1:N}) \to X_{1:N}$.
- pela desigualdade de processamento de dados em (A) teremos: $I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$.
- cadeia de Markov (B): $\theta \to X_{1:N} \to T(X_{1:N})$.
- pela desigualdade de processamento de dados em (B) teremos: $I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N}))$.
- então, (A) e (B) $\Rightarrow I(\theta; X_{1:N}) = I(\theta; T(X_{1:N}))$, e nenhuma informação é perdida sobre θ indo de $X_{1:N}$ para $T(X_{1:N})$.

Desigualdade de Processamento de Dados e Estatística II

Definição (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

$$I(\theta; X) = I(\theta; T(X)) \ \forall \theta \tag{163}$$

Uma estatística suficiente preserva a informação mútua e reciprocamente

$$X \perp\!\!\!\perp \theta | T(X) \tag{164}$$

- i.e., uma cadeira de Markov (A) é condição suficiente para suficiência de uma estatística.
- uma estatística suficiente é utilizada para estimar os parâmetros a partir dos dados: no limite em que temos infinitos dados, teremos uma estimativa exata (consistência assintótica).

Estatística Suficiente I

Exemplo

Seja $X_1, \ldots, X_N$, $X_i \in \{0, 1\}$, uma sequência i.i.d. de lances de uma moeda com parâmetro $\theta = Pr(X_i = 1)$. Dado N, o número de 1's é uma estatística suficiente para θ.

$$T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i \tag{165}$$

Dado T, todas as sequências com o mesmo número de 1's são igualmente prováveis e independentes do parâmetro θ.

...

Estatística Suficiente II

Exemplo

continuação...

Existem $\binom{N}{k}$ sequências de comprimento N com k 1's e são todas equiprováveis. $Pr(X_{1:N} = x_{1:N}) = \theta^k(1-\theta)^{N-k}$. Então

$$Pr\{(X_1, \ldots, X_N) = (x_1, \ldots, x_N) | \sum_{i=1}^{N} X_i = k\} = \begin{cases} \frac{1}{\binom{N}{k}} & \text{se } \sum_i x_i = k \\ 0 & \text{caso contrário} \end{cases} \quad (166)$$

Temos então que $\theta \to \sum X_i \to (X_1, \ldots, X_N)$ forma uma cadeia de Markov e T é uma estatística suficiente para θ (dado $\sum X_i$, a sequência $(X_1, \ldots, X_N)$ é estatisticamente independente de θ).

Teorema da Fatoração de Fisher-Neyman I

Teorema (Teorema da Fatoração de Fisher-Neyman)

Se a função densidade de probabilidade é $f_\theta(x)$, então T é suficiente para θ se e somente se podemos encontrar funções não-negativas g e h tais que

$$f_\theta(x) = h(x)g_\theta(T(x)), \tag{167}$$

i.e., a densidade f pode ser fatorada em um produto tal que um fator h não depende de θ e o outro fator, que depende de θ, dependerá de x apenas por meios de $T(X)$.

Estatística Suficiente I

Teorema (Estatística Suficiente)

$T(\cdot)$ é suficiente para θ sse a probabilidade $p(x_{1:N}|\theta)$ pode ser escrita como o produto

$$p(x_{1:N}|\theta) = g(T, \Theta)h(x_{1:N}) \tag{168}$$

$$\begin{aligned} p(x_{1:N}|\theta) &= g(T, \Theta)h(x_{1:N}) \\ &= g(T, \Theta)h(x_{1:N}, T(x_{1:N})) \end{aligned} \tag{169}$$

Estatística Suficiente II

Definição (Independencia Condicional)

Dadas três variáveis aleatórias A, B, C, temos que $A \perp\!\!\!\perp B|C$ sse existem funções g e h tais que $p(a, b, c)$ possa ser reescrita na forma

$$p(a, b, c) = g(a, c)h(b, c) \tag{170}$$

Estatística Suficiente I

Exemplo

Se X possui distribuição normal com média θ e variância 1

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = N(\theta, 1) \tag{171}$$

e $X_1, \ldots, X_n$ são tiradas de forma independente de acordo com esta distribuição. Uma estatística suficiente para θ é a média amostral

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{172}$$

...

Estatística Suficiente II

Exemplo

continuação...

$$
\begin{aligned}
f_\theta(x_1,\ldots,x_n) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} e^{\sum_{i=1}^n (x_i\theta-\theta^2/2)} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} e^{\theta n\left(\frac{1}{n}\sum_{i=1}^n x_i-\frac{\theta}{2}\right)} \\
&= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}}_{h(x_1,\ldots,x_n)} \underbrace{e^{\theta n\left(T(X_{1:n})-\frac{\theta}{2}\right)}}_{g_\theta(T(x_{1:n}))} \qquad (173)
\end{aligned}
$$

Então, pelo teorema de Fisher-Neyman, podemos concluir que a média amostral é uma estatística suficiente para θ quando X possui distribuição normal.

Tipo da Amostra I

Exemplo

Seja $X_1, \ldots, X_N \equiv X_{1:N}$ uma amostra de comprimento N de uma variável aleatória discreta D-ária. Então $x_i \in \mathcal{X}$, o tamanho do alfabeto é $D = |\mathcal{X}|$, e $\mathcal{X} = \{a_1, \ldots, a_D\}$.
Define-se uma estatística: o histograma empírico da amostra.

$$P_{x_{1:N}} \triangleq \left(\frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right), \tag{174}$$

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. É uma estatística suficiente?

...

Tipo da Amostra II

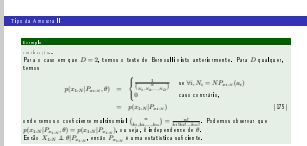
Exemplo

continuação...

Para o caso em que $D = 2$, temos o teste de Bernoulli visto anteriormente. Para D qualquer, temos

$$\begin{aligned} p(x_{1:N}|P_{x_{1:N}}, \theta) &= \begin{cases} \frac{1}{\binom{N}{N_1, N_2, \ldots, N_D}} & \text{se } \forall i, N_i = N P_{x_{1:N}}(a_i) \\ 0 & \text{caso contrário,} \end{cases} \\ &= p(x_{1:N}|P_{x_{1:N}}) \end{aligned} \tag{175}$$

onde temos o coeficiente multinomial $\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \ldots k_m!}$. 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.



Teorema Multinomial

$$(x_1 + x_2 + \ldots + x_m)^n = \sum_{k_1+k_2+\ldots+k_m=n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{1 \leq t \leq m} x_t^{k_t} \tag{176}$$

Caso Binário - Suficiência do Tipo I

Exemplo

- $X_i \in \{0,1\}$, $T(x_{1:N}) =$ número de 1s em $x_{1:N}$.
- A probabilidade conjunta:

$$p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in \mathcal{X}} p(a)^{N(a|x_{1:N})} = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{177}$$

- Evento $\{x_{1:N}, T(x_{1:n}) = k\}$ quando k é o verdadeiro número de 1s em $x_{1:N}$ e é o mesmo que o evento $\{x_{1:n}\}$. Quando k não é o número de 1s, temos probabilidade zero (impossível).

...

Caso Binário - Suficiência do Tipo II

Exemplo

continuação...

- Marginal $p(\theta, T(x_{1:N}) = k)$:

$$\begin{aligned} p(\theta, T(x_{1:N}) = k) &= \sum_{x_{1:N}} p(x_{1:N}, T(x_{1:N}) = k, \theta) \\ &= \sum_{x_{1:N}: T(x_{1:N}) = k} p(x_{1:N}, T(x_{1:N}) = k, \theta) \\ &= \binom{N}{k} p(0)^{N-k} p(1)^k \end{aligned} \tag{178}$$

...

Caso Binário - Suficiência do Tipo III

Exemplo

continuação...

- A probabilidade conjunta

$$p(x_{1:N}, T(x_{1:N}), \theta) = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{179}$$

- A marginal

$$p(\theta, T(x_{1:N}) = k) = \binom{N}{k} p(0)^{N-k} p(1)^k \tag{180}$$

- Então

$$p(x_{1:N}|T, \Theta) = \frac{p(x_{1:N}, T, \Theta)}{p(T, \Theta)} = \begin{cases} \frac{1}{\binom{N}{k}} & \text{se } \sum_i x_i = k \\ 0 & \text{caso contrário} \end{cases} \tag{181}$$

Estatística Mínima Suficiente I

Definição

Uma estatística $T(X)$ é uma estatística mínima suficiente em relação a $\{p_\theta(x)\}$ se ela for uma função de todas as demais estatísticas suficientes U.

- Sabemos pela definição de T mínima e qualquer outra estatística suficiente U que $\theta \to X_{1:N} \to U(X_{1:N}) \to T(X_{1:N})$
- Interpretando com relação à desigualdade do processamento de dados, temos

$$\theta \to T(X) \to U(X) \to X \tag{182}$$

- A estatística mínima suficiente T fornece qualquer outra estatística U independente do parâmetro θ.
- O fato de que é uma estatística significa que $p(X|T,U,\theta) = p(X|T,U) = p(X|T)$, o que significa que T é, para todos propósitos, uma substituto estatístico mínimo para θ no cálculo da probabilidade.

Estatística Suficiente I

Exemplo (Entropia Condicional Nula)

Mostre que se $H(Y|X) = 0$, então Y é uma função de X, i.e., para todo x com $p(x) > 0$, existe apenas um possível valor de y com $p(x, y) > 0$.

solução

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

$$p(y_1|x_0) = \frac{p(x_0, y_1)}{p(x_0)} \text{ e } p(y_2|x_0) = \frac{p(x_0, y_2)}{p(x_0)} \tag{183}$$

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

Estatística Suficiente II

Exemplo (Entropia Condicional Nula)

continuação...

$$
\begin{aligned}
H(Y|X) &= E[H(Y|X)] \\
&= -\sum_x p(x) H(Y|X=x) \\
&= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&\geq -p(x_0) \sum_y p(y|x_0) \log p(y|x_0) \\
&\geq -\underbrace{p(x_0)}_{>0}[\underbrace{p(y_1|x_0)}_{>0} \underbrace{\log p(y_1|x_0)}_{<0} + \underbrace{p(y_2|x_0)}_{>0} \underbrace{\log p(y_2|x_0)}_{<0}] \\
&> 0 \qquad (184)
\end{aligned}
$$

...

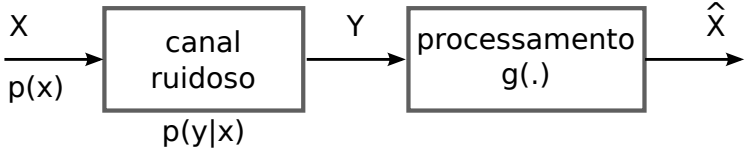
Estatística Suficiente III

Exemplo (Entropia Condicional Nula)

continuação...

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.

Erro nas Comunicações I



- $\hat{X}$ é uma estimativa de X.
- a estimativa é errada quando $X \neq \hat{X}$
- probabilidade de erro: $P_e \triangleq p(X \neq \hat{X})$
- podemos relacionar a entropia condicional $H(X|Y)$ com a probabilidade de erro P_e?
- sabemos (exercício anterior) que a entropia condicional $H(X|Y)$ é nula se e somente se X for uma função de Y
- esperamos ser capazes de estimar X com baixa probabilidade de erro apenas quando a entropia condicional $H(X|Y)$ for pequena

Desigualdade de Fano

Teorema (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

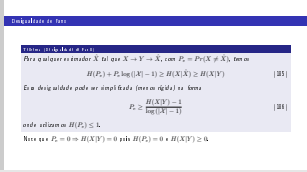
$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{185}$$

Esta desigualdade pode ser simplificada (menos rígida) na forma

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)} \tag{186}$$

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

Note que $P_e = 0 \Rightarrow H(X|Y) = 0$ pois $H(P_e) = 0$ e $H(X|Y) \geq 0$.



Esta desigualdade será utilizada para provar o reverso no teorema de codificação de Shannon, i.e., que qualquer código com probabilidade de erro $\to 0$, à medida que o comprimento do bloco cresce, devemos ter uma taxa $R < C$ (a capacidade do canal, a ser definida).
Para o caso de um alfabeto binário ($|\mathcal{X}| = 2$), a desigualdade de Fano na forma da Equação 186 não poderá ser aplicada. Devemos então utilizar a forma mais relaxada:

$$P_e \geq \frac{H(X|Y) - 1}{\log\left(|\mathcal{X}| - 1\right)} > \frac{H(X|Y) - 1}{\log|\mathcal{X}|} \tag{187}$$

Desigualdade de Fano I

Demonstração.

Definir uma função de erro:

$$E = \begin{cases} 1 & , \text{ se } \hat{X} \neq X (\text{erro}) \\ 0 & , \text{ se } \hat{X} = X (\text{sem erro}) \end{cases} \tag{188}$$

...

Desigualdade de Fano II

Demonstração.

continuação...

Utilizando a regra da cadeia temos:

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

ou

$$= \underbrace{H(E|\hat{X})}_{\leq H(E)=H(P_e)} + \underbrace{H(X|E,\hat{X})}_{\leq P_e \log(|\mathcal{X}|-1)} \quad (189)$$

- O erro é uma função determinística de X e $\hat{X}$, então, sabendo X e $\hat{X}$, determinamos E. Desta forma: $H(E|X,\hat{X}) = 0$.
- Condicionar só pode reduzir a entropia: $H(E|\hat{X}) \leq H(E) = H(P_e)$.
- Veremos abaixo que $H(X|E,\hat{X}) \leq P_e \log(|\mathcal{X}| - 1)$.

...

Desigualdade de Fano III

Demonstração.

continuação...

$$\begin{aligned} H(X|\hat{X},E) &= p(E=0)\underbrace{H(X|\hat{X},E=0)}_{=0} + p(E=1)H(X|\hat{X},E=1) \\ &= (1-P_e)0 + P_e H(X|\hat{X},E=1) \leq P_e \log(|\mathcal{X}|-1) \end{aligned}$$

- Se não há erro e conhecemos $\hat{X}$, então determinamos X. Não existe entropia residual em X quando é dado $\hat{X}$ e $E=0$. Logo $H(X|\hat{X},E=0)=0$.
- 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.

...

Desigualdade de Fano IV

Demonstração.

continuação...

Temos então

$$\begin{aligned} H(X|\hat{X}) &= H(E|\hat{X}) + H(X|E,\hat{X}) \\ &\leq H(P_e) + P_e \log(|\mathcal{X}| - 1) \end{aligned} \tag{190}$$

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

$$\begin{aligned} I(X;Y) &\geq I(X;\hat{X}) \\ H(X) - H(X|Y) &\geq H(X) - H(X|\hat{X}) \\ H(X|\hat{X}) &\geq H(X|Y) \end{aligned} \tag{191}$$

...

Desigualdade de Fano V

Demonstração.

continuação...

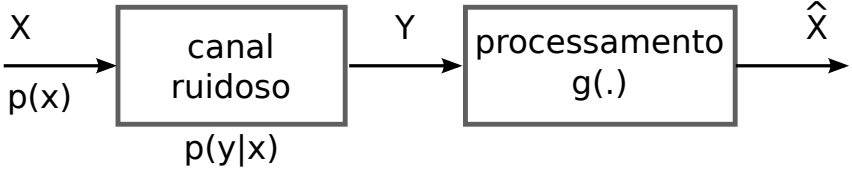
Então, utilizando as Equações 190 e 191, obtemos

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{192}$$

□

Desigualdade de Fano - Sumário

Considere a seguinte situação: enviamos X através de um canal ruidoso, recebemos Y e realizamos algum pós-processamento.



$\hat{X}$ é uma estimativa de X.

- Erro: $X \neq \hat{X}$; com probabilidade $P_e \triangleq p(X \neq \hat{X})$.
- Intuitivamente, a entropia condicional deveria nos dizer algo sobre a probabilidade de erro. Na verdade temos o seguinte:

Teorema (Desigualdade de Fano)

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{193}$$

Desigualdade de Fano I

Exemplo

Considere uma v.a. 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.

...

Desigualdade de Fano II

Exemplo

continuação...

solução

A distribuição condicional $p(y|x)$ é apresentada na tabela abaixo:

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	$1/2$	$1/2$

...

Desigualdade de Fano III

Exemplo

continuação...

A efetiva probabilidade de erro é dada por

$$\begin{aligned} P_e &= 1 - P_a \text{ (prob. de acerto)} \\ &= 1 - \sum_{i=1}^{5} P(x_i = y_i) \\ &= 1 - (p(y=1|x=1)p(x=1) + p(y=2|x=2)p(x=2) + 0 + 0 + 0) \\ &= 1 - \left(\frac{6}{7}0.35 + \frac{6}{7}0.35\right) = 0.4 = \frac{2}{5} \end{aligned} \tag{194}$$

...

Desigualdade de Fano IV

Exemplo

continuação...

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

$$P_e \geq \frac{H(X|Y) - 1}{\log\left(|\mathcal{X}| - 1\right)} \tag{195}$$

...

Desigualdade de Fano V

Exemplo

continuação...

Precisaremos calcular

$$\begin{aligned} H(X|Y) &= -\sum_{x,y} p(x,y)\log p(x|y) \\ &= -\sum_{x,y} p(y|x)p(x)\log p(x|y) \end{aligned} \tag{196}$$

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

...

Desigualdade de Fano VI

Exemplo

continuação...

$$
\begin{aligned}
P(X|Y=1) &= \frac{P(X,Y=1)}{P(Y=1)} = \frac{P(Y=1|X)P(X)}{P(Y=1)} \\
&= \frac{(\frac{6}{7},\frac{1}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2}) \cdot (0.35,0.35,0.1,0.1,0.1)}{\sum\left((\frac{6}{7},\frac{1}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2}) \cdot (0.35,0.35,0.1,0.1,0.1)\right)} \\
&= \frac{(0.3,0.05,0.05,0.05,0.05)}{1/2} \\
&= (0.6,0.1,0.1,0.1,0.1) \qquad (197)
\end{aligned}
$$

...

Desigualdade de Fano VII

Exemplo

continuação...

$$\begin{aligned} P(X|Y=2) &= \frac{(\frac{1}{7},\frac{6}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)}{\sum\left((\frac{1}{7},\frac{6}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)\right)} \\ &= (0.1,0.6,0.1,0.1,0.1) \end{aligned} \tag{198}$$

...

Desigualdade de Fano VIII

Exemplo

continuação...

Desta forma teremos

$$\begin{aligned} H(X|Y) &= H(X|Y=1)P(Y=1) + H(X|Y=2)P(Y=2) \\ &= -\frac{1}{2}\left(0.6\log 0.6 + 4\times 0.1\log 0.1\right) - \frac{1}{2}\left(4\times 0.1\log 0.1 + 0.6\log 0.6\right) \\ &= 1.771 \text{ bits.} \end{aligned} \tag{199}$$

Utilizando a desigualdade de Fano

$$P_e \geq \frac{H(X|Y)-1}{\log\left(|\mathcal{X}|-1\right)} = \frac{1.771-1}{\log(5-1)} = 0.3855 \tag{200}$$

Desigualdade de Fano I

Lema

Se X e X' são i.i.d. com entropia $H(X)$,

$$Pr(X = X') \geq 2^{-H(X)} \tag{201}$$

com igualdade se e somente se X possuir distribuição uniforme.

$$\begin{aligned} Pr(X = X') &= Pr(X = x_1|X' = x_1)Pr(X' = x_1) + \ldots + \\ &\quad Pr(X = x_n|X' = x_n)Pr(X' = x_n) \\ &= Pr(X = x_1)Pr(X' = x_1) + \ldots + \\ &\quad Pr(X = x_n)Pr(X' = x_n) \\ &= p^2(x_1) + \ldots + p^2(x_n) = \sum_x p^2(x) \end{aligned} \tag{202}$$

Desigualdade de Fano II

Demonstração.

Suponha que $X \sim p(x)$. Pela desigualdade de Jensen temos

$$2^{E[\log p(X)]} \leq E[2^{\log p(X)}] \tag{203}$$

pois 2^x é convexa. Logo,

$$\begin{aligned} 2^{-H(X)} &= 2^{\sum_x p(x)\log p(x)} = 2^{E[\log p(X)]} \\ &\leq E[2^{\log p(X)}] \\ &= \sum_x p(x)2^{\log p(x)} = \sum_x p(x)p(x) \\ &= \sum_x p^2(x) = Pr(X = X') \end{aligned} \tag{204}$$

□

Desigualdade de Fano III

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.

Desigualdade de Fano I

Corolário

Seja X, X' independentes com $X \sim p(x)$ e $X' \sim q(x)$, $x, x' \in \mathcal{X}$, então

$$\begin{aligned} Pr(X = X') &\geq 2^{-H(p)-D(p||q)} \\ Pr(X = X') &\geq 2^{-H(q)-D(q||p)} \end{aligned} \tag{205}$$

ou seja

$$Pr(X = X') \geq \max\left(2^{-H(p)-D(p||q)}, 2^{-H(q)-D(q||p)}\right) \tag{206}$$

Desigualdade de Fano II

Demonstração.

$$
\begin{aligned}
2^{-H(p)-D(p||q)} &= 2^{\sum_x p(x)\log p(x)+\sum_x p(x)\log \frac{q(x)}{p(x)}} \\
&= 2^{\sum_x p(x)\log q(x)} \\
&= 2^{E_p[\log q(X)]} \\
&\leq \sum_x p(x)2^{\log q(x)} \text{ (Jensen)} \\
&= \sum_x p(x)q(x) \\
&= Pr(X = X') \qquad (207)
\end{aligned}
$$

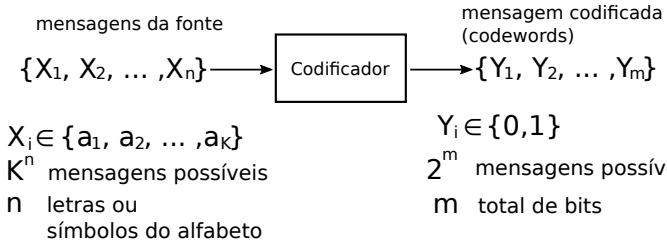
□

Propriedade da Equipartição Assintótica

- Vamos considerar blocos de realizações de uma variável aleatória (i.e., vetores aleatórios de comprimento n). $n =$ tamanho do bloco.
- Sejam $X_1, X_2, \ldots, X_n$ v.a.s i.i.d. com distribuição p (dizemos $X_i \sim p(x)$).
- Existem K símbolos possíveis (alfabeto ou espaço-estado de tamanho K), então $X_i \in \{a_1, a_2, \ldots, a_K\}$.
- Consideram n variáveis aleatórias $(X_1, X_2, \ldots, X_n)$, existem K^n possíveis realizações.

Propriedade da Equipartição Assintótica

Suponha que desejamos codificar as K^n possíveis realizações com uma sequência de dígitos binários de comprimento m. Então, existem 2^m palavras de código (*codewords*).



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

$$2^m \geq K^n \tag{208}$$

ou seja

$$m \geq (\log K)n \tag{209}$$

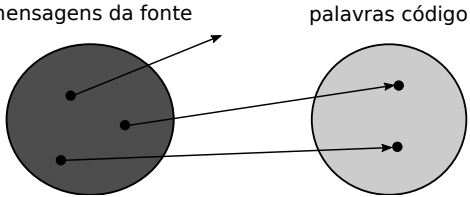
Propriedade da Equipartição Assintótica

- Quantos bits por letra da fonte utilizamos?

$$\text{taxa} = \frac{m}{n} \geq \log K \text{ bits por letra da fonte} \tag{210}$$

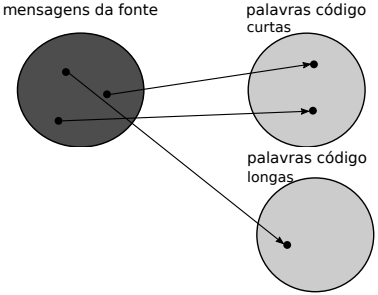
  Exemplo: 26 letras, precisaremos de $\lceil \log K \rceil = 5$bits.

- Podemos utilizar menos bits por símbolo emitido pela fonte (na média) e ainda sim não ter erro? Sim.
- Algumas mensagens da fonte poderiam não ter a elas um código associado.



Propriedade da Equipartição Assintótica

- Ao invés de descartar algumas mensagens, podemos associar a elas palavras longas e às outras palavras associamos palavras curtas.



- Em qualquer um dos casos, quando n é grande suficiente, podemos fazer com que a probabilidade, de se obter uma dessas mensagens da fonte que gerariam erro (ou que teriam palavras longas associadas), muito pequena.

Probabilidade de Palavras da Fonte

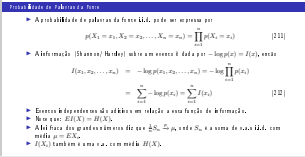
- A probabilidade de palavras da fonte i.i.d. pode ser expressa por

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \tag{211}$$

- A informação (Shannon/Hartley) sobre um evento é dada por $-\log p(x) = I(x)$, então

$$\begin{aligned} I(x_1, x_2, \ldots, x_n) &= -\log p(x_1, x_2, \ldots, x_n) = -\log \prod_{i=1}^{n} p(x_i) \\ &= \sum_{i=1}^{n} -\log p(x_i) = \sum_{i=1}^{n} I(x_i) \end{aligned} \tag{212}$$

- Eventos independentes são aditivos em relação a esta função de informação.
- Note que: $EI(X) = H(X)$.
- A lei fraca dos grandes números 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 $\mu = EX_i$.
- $I(X_i)$ também é uma v.a. com média $H(X)$.


Probabilidade de Palavras da Fonte

- A probabilidade de palavras da fonte i.i.d. pode ser expressa por

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \quad (211)$$

- A informação (Shannon/Hartley) sobre um evento é dada por $-\log p(x) = I(x)$, então

$$\begin{aligned} I(x_1, x_2, \ldots, x_n) &= -\log p(x_1, x_2, \ldots, x_n) = -\log \prod_{i=1}^{n} p(x_i) \\ &= \sum_{i=1}^{n} -\log p(x_i) = \sum_{i=1}^{n} I(x_i) \end{aligned} \quad (212)$$

- Entropia é definida em função de atributos em relação a sua função de informação.
- Note que: $EI(X) = H(X)$.
- Abaixo fora dos grandes números diz que $\frac{1}{n} S_n \xrightarrow{n} \mu$, onde S_n é a soma de n.a.s i.i.d. com média $\mu = EX_i$.
- $I(X_i)$ também é uma v.a., com média $H(X)$.

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,

$$\lim_{n \to \infty} P\left(|\overline{X_n} - \mu| < \varepsilon\right) = 1. \quad (213)$$

Lei fraca dos grandes números e Entropia

- Combinando o que vimos anteriormente, obtemos

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i) \xrightarrow[n\to\infty]{p} H(X) \tag{214}$$

- Se n fica grande suficiente, obteremos

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} I(x_i) &\approx H(X) \text{ onde } \forall i, x_i \sim p(x) \\
-\frac{1}{n}\sum_{i=1}^{n} \log p(x_i) &\approx H(X) \\
-\log \prod_{i=1}^{n} p(x_i) &\approx nH(X) \\
-\log p(x_1, x_2, \ldots, x_n) &\approx nH(X) \\
p(x_1, x_2, \ldots, x_n) &\approx 2^{-nH(X)}
\end{aligned} \tag{215}$$

Propriedade da Equipartição Assintótica

Quando n é grande suficiente, teremos

$$p(x_1, x_2, \ldots, x_n) \approx 2^{-nH(X)} \tag{216}$$

- Esta probabilidade não depende da sequência em si. Depende apenas do comprimento n e da entropia da v.a.
- 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.

Quase todos eventos são quase equiprováveis

- Se $X_1, X_2, \ldots, X_n$ são i.i.d. e $X_i \sim p(x)$ para todo i, e se n é grande suficiente, então qualquer amostra $x_1, x_2, \ldots, x_n$ terá probabilidade da amostra essencialmente independente da amostra, i.e.,

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \tag{217}$$

  onde $H(X)$ é a entropia de $p(x)$.
- Então, podem existir no máximo 2^{nH} amostras, e pode ser que $2^{nH} \ll K^n$.
- Estas amostras que ocorrem são chamadas de típicas, e são representadas por $A_\epsilon^{(n)}$.
- Uma grande porção de $\mathcal{X}^n$ não irá ocorrer, i.e., pode acontecer que $2^{nH} \ll |\mathcal{X}^n| = K^n$.

Conjunto Típico

- Seja $A_\epsilon^{(n)}$ o conjunto das sequências típicas (i.e., aquelas com probabilidade 2^{-nH}).
- Se "todos" eventos possuem a mesma probabilidade p, então existem $1/p$ deles.
- O número de sequências típicas é

$$|A_\epsilon^{(n)}| \approx 2^{nH(X)}. \tag{218}$$

- Desta forma, para representar (ou codificar) as sequencias típicas, precisaremos de $nH(X)$ bits. Teremos então

$$m = nH(X) \tag{219}$$

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

Codificando apenas o Conjunto Típico

- Tomando $m = nH$, teremos que o número médio de bits por letra do alfabeto da fonte será dado por

$$\frac{m}{n} = H \text{ que pode ser } \leq \log K \tag{220}$$

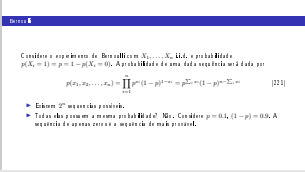
- Interpretações para a Entropia na codificação de fonte:
  1) A probabilidade de uma sequência típica é $2^{-nH(X)}$.
  2) O número de sequências típicas é $2^{nH(X)}$.
  3) O número de bits por símbolo da fonte é $H(X)$, quando codificamos apenas o conjunto típico.

Bernoulli

Considere o experimento de Bernoulli com $X_1, \ldots, X_n$ i.i.d. e probabilidade $p(X_i = 1) = p = 1 - p(X_i = 0)$. A probabilidade de uma dada sequência será dada por

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \tag{221}$$

- Existem 2^n sequencias possíveis.
- Todas elas possuem a mesma probabilidade? Não. Considere $p = 0.1$, $(1-p) = 0.9$. A sequência de apenas zeros é a sequência de mais provável.



Todas as sequências que possuem 'alguma' probabilidade, terão a mesma probabilidade? Depende do que queremos dizer com 'alguma'. Para valores pequenos de n, não, mas à medida que n cresce, algo acontece e a resposta 'sim' começa a ser a mais apropriada.

Propriedade de Equipartição Assintótica

- É possível prever a probabilidade de que uma determinada sequências terá uma probabilidade particular?

$$\Pr(p(X_1, X_2, \ldots, X_n) = \alpha) = ? \tag{222}$$

- Note que $p(X_1, X_2, \ldots, X_n)$ é uma variável aleatória. É uma probabilidade que é uma função do conjunto de variáveis aleatórias.
- Teremos

$$\Pr(p(X_1, X_2, \ldots, X_n) \approx 2^{-nH}) \approx 1 \tag{223}$$

quando n é grande suficiente.

- Quase todos os eventos (que ocorrem com alguma probabilidade) são todos equiprováveis.

Ensaio de Bernoulli

Exemplo

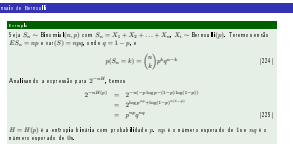
Seja $S_n \sim$ Binomial(n,p) com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim$ Bernoulli(p). Teremos então $ES_n = np$ e var$(S) = npq$, onde $q = 1-p$, e

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{224}$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n(-p\log p-(1-p)\log(1-p))} \\ &= 2^{\log p^{np}+\log(1-p)^{n(1-p)}} \\ &= p^{np}q^{nq} \end{aligned} \tag{225}$$

$H = H(p)$ é a entropia binária com probabilidade p. np é o número esperado de 1s e nq é o número esperado de 0s.


Ensaio de Bernoulli

Exemplo

Seja $S_n \sim \text{Binomial}(n, p)$ com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim \text{Bernoulli}(p)$. Temos ainda $E[S_n] = np$ e $\text{var}(S) = np(q)$, onde $q = 1 - p$, e

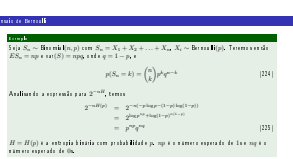
$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (224)$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n[-p\log_2 p - (1-p)\log_2(1-p)]} \\ &= 2^{np\log_2 p^{-1} + n\log_2(1-p)^{n(1-p)}} \\ &= p^{np} q^{nq} \end{aligned} \quad (225)$$

$H = H(p)$ é a entropia binária com probabilidade p, np é o número esperado de 1s e nq é o número esperado de 0s.

- Todas as sequências que ocorrem são aquelas cujo número de 1s e 0s são aproximadamente iguais aos seus valores esperados.
- Nenhum outra sequência possui probabilidade significativa.
- A sequência $X_1, X_2, \ldots, X_n$ foi assumida como sendo i.i.d., entretanto podemos estender para cadeias de Markov e processos aleatórios estacionários ergódicos.


Ensaio de Bernoulli

Exemplo

Seja $S_n \sim \text{Binomial}(n,p)$ com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim \text{Bernoulli}(p)$. Teremos ainda $E[S_n] = np$ e $\text{var}[S] = npq$, onde $q = 1 - p$, e

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (224)$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n[-p\log_2 p - (1-p)\log_2(1-p)]} \\ &= 2^{\log_2 p^{np}} + \log_2(1-p)^{n(1-p)} \\ &= p^{np} q^{nq} \end{aligned} \quad (225)$$

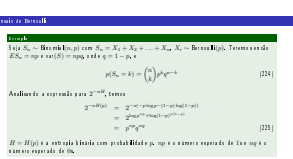
$H = H(p)$ é a entropia binária com probabilidade p, sendo np o número esperado de 1s e nq o número esperado de 0s.

O cálculo computacional do coeficiente binomial pode apresentar perda de precisão numérica, por envolver fatoriais e frações de números muito grandes. Exemplo de execussão no GNU-Octave para calcular $\binom{100}{20}$:

```
» nchoosek(100,20)
warning: nchoosek: possible loss of precision
warning: called from
```

Uma possível solução é utilizar o pacote simbólico para efetuar os cálculos. Podemos verificar que realmente ocorreu erro de precisão numérica ao realizar o cálculo computacional.

```
» double(nchoosek(sym(100),sym(20))) - nchoosek(100,20)
ans = -65536
```


Ensaio de Bernoulli

Exemplo

Seja $S_n \sim \text{Binomial}(n,p)$ com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim \text{Bernoulli}(p)$. Teremos ainda $E[S_n] = np$ e $\text{var}(S) = np(q)$, onde $q = 1-p$, e

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \quad (224)$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n[-p\log p-(1-p)\log(1-p)]} \\ &= 2^{\log p^{np}+\log(1-p)^{n(1-p)}} \\ &= p^{np}q^{nq} \end{aligned} \quad (225)$$

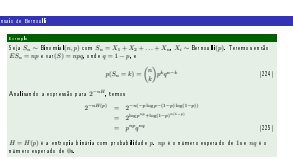
$H = H(p)$ é a entropia binária com probabilidade p, sendo np o número esperado de 1s e nq o número esperado de 0s.

Outra alternativa é realizar uma aproximação para calcular $\binom{N}{k}$.
Iremos utilizar a aproximação de Stirling para a função fatorial:

$$\ln n! = n\ln n - n + \mathcal{O}(\ln n) \quad (226)$$

Utilizando esta aproximação em $\ln\binom{N}{k}$, teremos

$$\begin{aligned} \ln\binom{N}{k} &\equiv \ln\frac{N!}{(N-k)!k!} = \ln N! - \ln(N-k)! - \ln k! \\ &\simeq N\ln N - N - (N-k)\ln(N-k) + (N-k) - k\ln k + k \\ &= \underbrace{(N-k)\ln N - (N-k)\ln N}_{=0} + N\ln N - N - (N-k)\ln(N-k) + (N-k) - k\ln k + k \\ &= (N-k)\ln\frac{N}{N-k} + k\ln\frac{N}{k} \\ &= N\left(-\frac{N-k}{N}\ln\frac{N-k}{k} - \frac{k}{N}\ln\frac{k}{N}\right) = NH_e\left(\frac{k}{N}\right). \end{aligned} \quad (227)$$


Ensaio de Bernoulli

Exemplo

Seja $S_n \sim \text{Binomial}(n,p)$ com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim \text{Bernoulli}(p)$. Teremos ainda $E[S_n] = np$ e $\text{var}[S] = np(q)$, onde $q = 1 - p$, e

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{224}$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n[-p\log_2 p - (1-p)\log_2(1-p)]} \\ &= 2^{\log_2 p^{np} + \log_2(1-p)^{n(1-p)}} \\ &= p^{np} q^{nq} \end{aligned} \tag{225}$$

$H = H(p)$ é a entropia binária com probabilidade p, np é o número esperado de 1s e nq é o número esperado de 0s.

Concluímos então que

$$\ln \binom{N}{k} \simeq N H_e \left(\frac{k}{N} \right) \tag{228}$$

e, como os temos em ambos os lados envolvem logaritmos, podemos realizar a mudança de base em ambos os lados (basta multiplicar por $\ln 2$),

$$\log \binom{N}{k} \simeq N H \left(\frac{k}{N} \right), \tag{229}$$

onde agora utilizamos a entropia binária em bits. Assim, teremos

$$\binom{N}{k} \simeq 2^{N H\left(\frac{k}{N}\right)}. \tag{230}$$

Propriedade da Equipartição Assintótica

Teorema (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

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X) \tag{231}$$

Demonstração.

$$\begin{aligned} -\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\log \prod_{i=1}^{n} p(X_i) \\ &= -\frac{1}{n}\sum_i \log p(X_i) \xrightarrow{p} E\log p(X) \\ &\quad \text{onde utilizamos a lei fraca dos números grandes} \\ &= H(X) \end{aligned} \tag{232}$$

Conjunto Típico

Definição (Conjunto Típico)

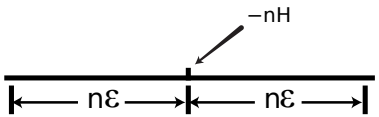
Um conjunto típico $A_\epsilon^{(n)}$ em relação a $p(x)$ é o conjunto de sequências $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ com propriedade

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{233}$$

De forma equivalente, podemos escrever

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \ldots, x_n) : | -\frac{1}{n} \log p(x_1, \ldots, x_n) - H| < \epsilon \right\} \tag{234}$$

O conjunto típico é formado pelas sequências com $\log$ da probabilidade dentro de seguinte extensão



Conjunto Típico

- O tamanho do conjunto típico de sequências produzidas pela fonte é tipicamente muito menor que o tamanho do conjunto de todas as sequências produzidas pela fonte.
- Uma sequencia típica não precisa ter probabilidade próxima daquela que é a sequência mais provável.
- Geralmente a sequência mais provável não está no conjunto típico.

Propriedades do Conjunto Típico

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

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

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \tag{235}$$

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.*

- O conjunto típico possui, essencialmente, probabilidade 1 (algo típico irá tipicamente ocorrer).
- Todos os itens neste conjunto terão a mesma probabilidade $\approx 2^{-nH}$.

Exemplo $K = |\{0, 1\}|$: Conjunto Típico

- Suponha o Ensaio de Bernoulli com distribuição uniforme, $K = 2$, $\mathcal{X} = \{0, 1\}$, $p = 0.5$, entropia $H = 1$ e $|A_\epsilon^{(n)}| = 2^{nH} = 2^n = K^n$, então todas as sequências ocorreram com igual probabilidade.
- Considere agora uma distribuição não-uniforme, $p = 0.1$ e $q = 1 - p = 0.9$, a entropia será $H \approx 0.469$. Considere $n = 100$, então $K^{100} = 2^{100} = 10^{\log_{10} 2^{100}} \approx 10^{100 \times 0.30103} \approx 10^{30}$, a capacidade representacional das sequencias da fonte. Mas $|A_\epsilon^{(n)}| = 2^{nH} = 10^{nH \times \log_{10} 2} \approx 10^{14} \ll 10^{30} \approx K^{100}$. O número de sequências típicas é muito menor que o número de possíveis sequências.
- Ineficiência: capacidade representacional é muito maior do que as coisas que ocorrem. O alfabeto da fonte é pobre para realizar compressão.
- Assuma ϵ muito pequeno, então onde foi parar a massa das $\approx 10^{30} - 10^{14}$ sequencias? (veremos adiante)

Conjunto Típicos são Típicos

- Pela definição
$$p(A_\epsilon^{(n)}) > 1 - \epsilon \text{ para qualquer } \epsilon > 0 \quad (236)$$
- Então $A_\epsilon^{(n)}$ possui praticamente toda probabilidade, e cada elemento em $A_\epsilon^{(n)}$ possui a mesma probabilidade, então
$$p(x) \approx 2^{-nH} \forall x \in A_\epsilon^{(n)} \quad (237)$$
- Exemplo: Ensaio de Bernoulli $X_i \sim \text{Bernoulli}(p)$ com $p(X_i = 1) = p = 1 - p(X_i = 0)$ e $p > 0.5$.
- Probabilidade de n 1s sucessivos é p^n e esta é a sequência mais provável.
- Probabilidade de uma sequência típica é 2^{-nH}.
- Para $n = 100$, $p - 0.9 = 1 - q$, a sequência mais provável possui probabilidade $p^n \approx 2.66 \times 10^{-5}$, mas uma sequencia típica possui probabilidade $2^{-nH} \approx 7.62 \times 10^{-15}$.

Exemplo $K = |\{0, 1\}|$: Conjunto Típico

- Suponha o Ensaio de Bernoulli com distribuição uniforme, $K = 2$, $\mathcal{X} = \{0, 1\}$, $p = 0.5$, entropia $H = 1$ e $|A_\epsilon^{(n)}| = 2^{nH} = 2^n = K^n$, então todas as sequências ocorreram com igual probabilidade.
- Considere agora uma distribuição não-uniforme, $p = 0.1$ e $q = 1 - p = 0.9$, a entropia será $H \approx 0.469$. Considere $n = 100$, então $K^{100} = 2^{100} = 10^{\log_{10} 2^{100}} \approx 10^{100 \times 0.30103} \approx 10^{30}$, a capacidade representacional das sequencias da fonte. Mas $|A_\epsilon^{(n)}| = 2^{nH} = 10^{nH \times \log_{10} 2} \approx 10^{14} \ll 10^{30} \approx K^{100}$. O número de sequências típicas é muito menor que o número de possíveis sequências.
- Ineficiência: capacidade representacional é muito maior do que as coisas que ocorrem. O alfabeto da fonte é pobre para realizar compressão.
- Assuma ϵ muito pequeno, então onde foi parar a massa das $\approx 10^{30} - 10^{14}$ sequencias? (veremos adiante)

Conjunto Típicos são Típicos

- Pela definição
$$p(A_\epsilon^{(n)}) > 1 - \epsilon \text{ para qualquer } \epsilon > 0 \quad (238)$$
- Então $A_\epsilon^{(n)}$ possui praticamente toda probabilidade, e cada elemento em $A_\epsilon^{(n)}$ possui a mesma probabilidade, então
$$p(x) \approx 2^{-nH} \forall x \in A_\epsilon^{(n)} \quad (239)$$
- Exemplo: Ensaio de Bernoulli $X_i \sim \text{Bernoulli}(p)$ com $p(X_i = 1) = p = 1 - p(X_i = 0)$ e $p > 0.5$.
- Probabilidade de n 1s sucessivos é p^n e esta é a sequência mais provável.
- Probabilidade de uma sequência típica é 2^{-nH}.
- Para $n = 100$, $p - 0.9 = 1 - q$, a sequência mais provável possui probabilidade $p^n \approx 2.66 \times 10^{-5}$, mas uma sequencia típica possui probabilidade $2^{-nH} \approx 7.62 \times 10^{-15}$.

Sequencias Não-Típicas não são típicas

- $p^n \gg 2^{-nH}$: a sequência mais provável é muito mais provável do que uma sequencia típica.
- O que ocorre com a probabilidade da sequencia mais provável, na média (por símbolo), quando n cresce

$$-\frac{1}{n}\log p^n = -\log p = \log 1/p \xrightarrow[n\to\infty]{p} H? \tag{240}$$

- O conjunto típico possui, essencialmente, toda a probabilidade $p(A_\epsilon^{(n)}) > 1-\epsilon$.
- A sequencia mais provável está no conjunto típico? Não, já que a probabilidade da sequencia mais provável não é 2^{-nH}.
- $n = 100$, $p = 0.9 = 1 - q$. Considere uma sequência com noventa 1s e dez 0s. A probabilidade é $p^{90}(1-p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH}$.
- Esta sequencia muito improvável é típica.

Probabilidade Média de Sequencias

- Pela PEA (Propriedade da Equipartição Assintótica), temos que se $(x_1, \ldots, x_n)$ é típico (i.e., $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$, para qualquer $\epsilon > 0$), então

$$-\frac{1}{n} \log p(x_1, \ldots, x_n) \xrightarrow[n\to\infty]{p} H \tag{241}$$

- Para as sequencias mais prováveis, quando n é grande suficiente, teremos (para $p > 0.5$)

$$-\frac{1}{n} \log p^n = -\log p = \log 1/p \xrightarrow[n\to\infty]{p} -\log p \tag{242}$$

- A probabilidade média da sequencia mais provável é bem diferente da sequencia típica.

Sequencias Típicas

- o conjunto típico possui essencialmente toda da probabilidade.
- Como pode uma sequencia com a maior probabilidade não ser típica, mas uma sequência com probabilidade muito menor ser típica?
- Existe um número exponencialmente crescente de sequencias típicas, cada uma com probabilidade menor do que a sequencia de maior probabilidade.
- A probabilidade individual de cada sequencia vai a zero quando n cresce.
- O tamanho do conjunto típico cresce rapidamente, quando $n \to \infty$, de forma que a probabilidade de $A_\epsilon^{(n)}$ vai a 1.
- O tamanho do conjunto das sequencias muito prováveis cresce lentamente, de forma que a probabilidade do conjunto vai a zero quando $n \to \infty$.

Sequencias Típicas I

A probabilidade de uma sequência $x_{1:n}$, onde $x_i \in \mathcal{X} = \{a_1, \ldots, a_K\}$, é dada por

$$P(x_{1:n}) = P(x_1)P(x_2)\ldots P(x_n) \approx p_1^{p_1 n} p_2^{p_2 n} \ldots p_K^{p_K n}, \tag{243}$$

onde consideramos que, em uma sequência muito longa, esperamos observar $p_1 n$ ocorrências do símbolo a_1, $p_2 n$ ocorrências do símbolo a_2, etc.
A informação associada a esta sequência típica é

$$\begin{aligned} \log \frac{1}{P(x_{1:n})} &\approx \sum_{i=1}^{K} \log p_i^{-p_i n} \\ &= n \sum_{i=1}^{K} p_i \log \frac{1}{p_i} = nH(X). \end{aligned} \tag{244}$$

Uma sequência típica terá probabilidade próxima de $2^{-nH(X)}$.

Sequencias Típicas II

x	$\log_2(P(x))$
...1.....1.....1....1.1.....1.....1.....1.....1.....11..	−50.1
.....1.....1.....1.....1...1.....1.....1....	−37.3
.....1...1..1..1....11..1.1.....11.....1..1.1..1..1.....1	−65.9
1.1..1.....1.....11.1..1.....1.....1....1..1.11.....	−56.4
...11.....1...1.....1.1.....1.....1...1...1.....1.....	−53.2
.....1.....1.....1.1.....1.....1.....1..1.....1.....	−43.7
.....1.....1.....1..1.....1.....1.....1....1.11.....	−46.8
.....1..1..1.....111.....1.....1.....1.1..1..1.....1	−56.4
.....1.....1.....1.....1.....1...1.....1.....	−37.3
.....1.....1.....1.....1.....1..1.1.1.1..1.....1	−43.7
1.....1.....1...1.....1.....1...1.....1.11..1.1..1.....	−56.4
.....11.1.....1.....1.....1.....1.....	−37.3
.1.....1...1.1.....1.....11.....1.1..1.....1.....11.....	−56.4
.....1..1..1.....1..11.1.1.1..1.....1.....1.....1.1.....	−59.5
.....11.1.....1...1..1.....1.....1.....1.....1.....	−46.8
.....	−15.2
11111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111	−332.1

Figura 5: 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).

Sequencias Típicas III

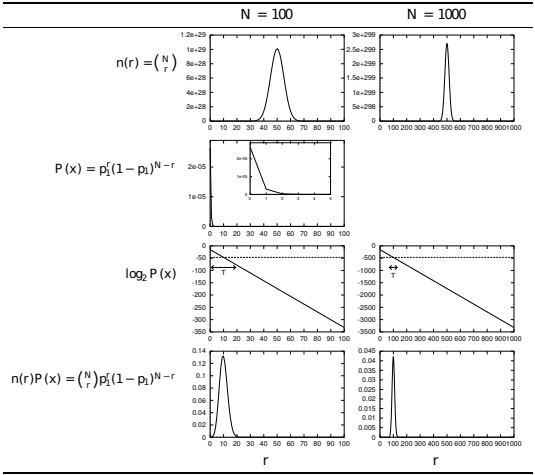


Figura 6: Para $p = 0.1$, $n = 100$ e $n = 1000$ os gráficos ilustram $n(r)$, o número de strings contendo r 1s; a probabilidade $P(x_{1:n})$ para uma string contendo r 1s; a mesma probabilidade em escala logarítmica; e a probabilidade total $n(r)P(x_{1:n})$ de todas as strings contendo r 1s (MacKay, 2003).

Sequencias Típicas IV

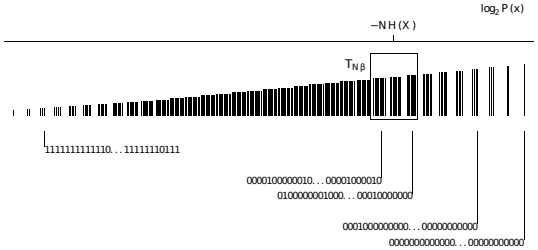


Figura 7: Diagrama esquemático ilustrando todas as sequencias no conjunto $\mathcal{X}^n$ ordenadas pela probabilidade (MacKay, 2003).

O termo **equipartição** é utilizado para descrever a ideia de que os membros do conjunto típico possuem aproximadamente a mesma probabilidade.

Sequencias Típica - Exemplo: distribuição Binomial I

$p(S_n = k) = \binom{n}{k} p^k q^{n-k}$, $S_n = X_1 + \ldots + X_n$, $X_i \sim$ Bernoulli(p) Abaixo o gráfico dos valores normalizados $S_n/n = k/n$ para $p = 0.5$.

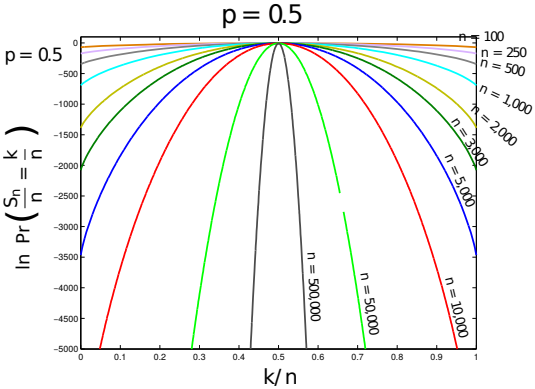


Figura 8: (Bilmes, 2013).

Sequencias Típica - Exemplo: distribuição Binomial II

$p(S_n = k) = \binom{n}{k} p^k q^{n-k}$, $S_n = X_1 + \ldots + X_n$, $X_i \sim$ Bernoulli(p) Abaixo o gráfico dos valores normalizados $S_n/n = k/n$ para $p = 0.9$.

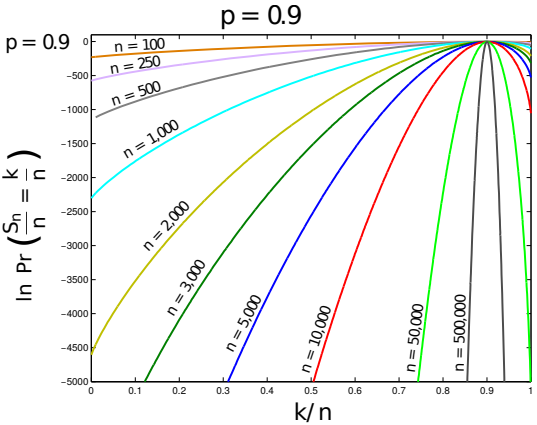
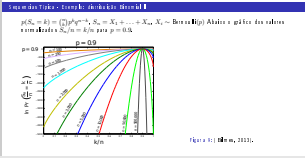


Figura 9: (Bilmes, 2013).



Para comparação, o número de átomos no universo observável é $\approx e^{187} \approx 10^{81}$.

Uma imagem full HD (1080×720 pixels) com 16 milhões de cores (24 bits por pixel) é representada por $K = 1080 \times 720 \times 24 = 18.662.400 \approx 10^7$ bits. Existem 2^K imagens possíveis, ou seja, $2^{18.662.400} \approx 10^{5.617 \times 10^6}$ sequências binárias que representam imagens full HD com 16 milhões de cores, ou seja, um número **muito** maior do que o número de átomos no universo. Obviamente, o conjunto de imagens que efetivamente ocorrem (ocorreram e ocorrerão ao longo de toda história da humanidade) é muito menos do que o número de possíveis imagens.

Propriedades de $A_\epsilon^{(n)}$

Teorema (Propridades de $A_\epsilon^{(n)}$)

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

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, \ldots, x_n) \leq H(X) + \epsilon \tag{245}$$

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.*

- O conjunto típico possui, essencialmente, probabilidade 1 (algo típico irá tipicamente ocorrer).
- Todos os itens neste conjunto terão a mesma probabilidade $\approx 2^{-nH}$.
- O número de elementos neste conjunto é $\approx 2^{nH}$.

Propriedades de $A_\epsilon^{(n)}$

Demonstração.

1) O primeiro apenas é uma reformulação da definição de PEA.
2) Vamos utilizar a definição expandida de convergência em probabilidade, dada na equação 231.

$$p(A_\epsilon^{(n)}) = p\left(|-\frac{1}{n}\sum_i \log p(x_i) - H| < \epsilon\right) > 1-\delta \tag{246}$$

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

$$p(A_\epsilon^{(n)}) > 1-\epsilon, \qquad \text{para } n \text{ grande suficiente } \forall\epsilon \tag{247}$$

□

Propriedades de $A_\epsilon^{(n)}$

Demonstração.

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

$$\begin{aligned} 1 &= \sum_x p(x) \geq \sum_{x \in A_\epsilon^{(n)}} p(x) \geq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)+\epsilon)} \\ &= |A_\epsilon^{(n)}| 2^{-n(H(X)+\epsilon)} \end{aligned} \tag{248}$$

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

□

Propriedades de $A_\epsilon^{(n)}$

Demonstração.

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

$$\begin{aligned} 1-\epsilon &< p(A_\epsilon^{(n)}) \leq \sum_{x \in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} \\ &= 2^{-n(H(X)-\epsilon)} |A_\epsilon^{(n)}| \end{aligned} \tag{249}$$

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

□

Codificação com Conjunto Típico

- Existe um código que pode alcançar uma taxa de compressão de $H(X) + \epsilon'$ bits por símbolo para qualquer $\epsilon' > 0$ enquanto o comprimento do bloco n (o número de símbolos da fonte que são codificados simultâneamente) for grande suficiente.
- Mesmo que as mensagens da fonte sejam consideradas i.i.d., é necessário codificá-las conjuntamente de forma a alcançar esta taxa.
- Fazemos isso utilizando $n(H + \epsilon) + 2$ bits para cada sequencia típica, e utilizamos $n \log K + 2$ bits para cada sequência atípica. Este código é $1 - 1$ e garantidamente livre de erros.
- De forma alternativa, podemos projetar para que exista um erro quando recebermos uma sequência atípica. Este código terá probabilidade de erro P_e.
- Em qualquer um dos casos, o comprimento esperado é

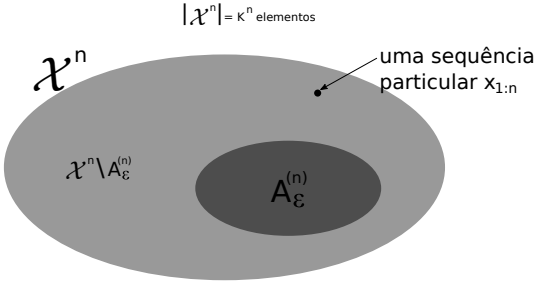
$$E[\frac{1}{n} l(X_{1:n})] \leq H(X) + \epsilon \tag{250}$$

- No segundo caso, $P_e \to 0$ quando $n \to \infty$.

Compressão de Dados até a Entropia da Fonte

Uma consequência importante é que podemos comprimir os dados até o limite da entropia da fonte.

- Ideia: considere $X_1, X_2, \ldots, X_n$ i.i.d. e $X_i \sim p(x)$.
- Particionar o conjunto de sequencias em dois blocos:
  - Conjunto típico $A_\epsilon^{(n)}$
  - e seu complemento, o conjunto não típico: $A_\epsilon^{(n)c} \triangleq \mathcal{X}^n \backslash A_\epsilon^{(n)}$.
- Partições, i.e., $A_\epsilon^{(n)} \cap A_\epsilon^{(n)c} = \emptyset$ e $A_\epsilon^{(n)} \cup A_\epsilon^{(n)c} = \mathcal{X}^n$



Compressão do Conjunto Típico I

- Vamos 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

$$\lceil n(H+\epsilon) \rceil \leq n(H+\epsilon) + 1 \text{ bits.} \tag{251}$$

- Podemos utilizar um bit extra para indicar se o elemento está no conjunto típico ou não, i.e., vamos utilizar

$$(b_0, b_1, b_2, \ldots, b_{\lceil n(H+\epsilon) \rceil}) \tag{252}$$

  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 é $\leq n(H+\epsilon) + 2$.
- Para os elementos do conjunto não típico, vamos utilizar $\lceil \log |\mathcal{X}|^n \rceil \leq n \log K + 1$ bits.

Compressão do Conjunto Típico II

- Vamos utilizar um vetor binário da forma

$$(b_0, b_1, b_2, \ldots, b_{\lceil \log |\mathcal{X}|^n \rceil}) \tag{253}$$

  onde $b_0 = 1$, indicando a atipicidade.
- O número total de bits para uma sequência atípica é $\leq n \log K + 2$.
- Este código criado é 1-pra-1, sendo fácil codificar e decodificar, dado o *codebook* (mapeamento).
- Para o conjunto não-típico $A_\epsilon^{(n)c}$ estamos utilizando mais bits do que o necessário, já que

$$|A_\epsilon^{(n)c}| = |\mathcal{X}^n| - |A_\epsilon^{(n)}| = K^n - |A_\epsilon^{(n)}| \leq K^n, \tag{254}$$

  mas isto não importará, como veremos adiante.
- As sequências típicas possuem um comprimento descritivo curto, $\approx nH$.
- Seja $l(x_{1:n})$ o comprimento da palavra (*codeword*) associada à sequência $x_{1:n}$.

Compressão do Conjunto Típico III

- $l(X_{1:n})$ 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.

Comprimento Esperado I

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

$$\begin{aligned}
El(X_{1:n}) &= \sum_{x_{1:n}} p(x_{1:n}) l(x_{1:n}) \\
&= \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) l(x_{1:n}) + \sum_{x_{1:n} \in A_\epsilon^{(n)c}} p(x_{1:n}) l(x_{1:n}) \\
&\leq \sum_{x_{1:n} \in A_\epsilon^{(n)}} p(x_{1:n}) [n(H+\epsilon) + 2] + \sum_{x_{1:n} \in A_\epsilon^{(n)c}} p(x_{1:n}) [n \log K + 2] \\
&= \underbrace{p(A_\epsilon^{(n)})}_{\leq 1} [n(H+\epsilon) + 2] + \underbrace{p(A_\epsilon^{(n)c})}_{< \epsilon} [n \log K + 2] \\
&\leq n(H+\epsilon) + 2 + \epsilon n \log K + \epsilon 2 \\
&= n[H + \underbrace{\epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n}}_{\epsilon'}] = n(H + \epsilon') \qquad (255)
\end{aligned}$$

Comprimento Esperado II

- Podemos fazer ϵ' tão pequeno quanto queremos, fazendo ϵ pequeno e n grande.

$$\epsilon' = \epsilon + \epsilon \log K + \frac{2}{n} + \frac{2\epsilon}{n} \tag{256}$$

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

Teorema (Primeiro Teorema de Shannon)

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

$$E[\frac{1}{n} l(X_{1:n})] \leq H(X) + \epsilon \tag{257}$$

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

Comprimento Esperado III

- Será necessário no máximo $nH(X)$ bits para representar $X_{1:n}$, na média, ou $H(X)$ bits por símbolo do alfabeto da fonte.
- O primeiro teorema de Shannon diz que é possível (utilizando blocos de comprimento longo) comprimir até o limite da entropia.
- Exemplo: *online coding* - codificar apenas o que é encontrado, sabendo que, para n grande suficiente, aquilo que encontrar deve ser típico.
- É necessário ainda mostrar que não é possível comprimir abaixo do valor da entropia sem causar erros.

Existem outro conjunto muito provável? I

- Sabemos que $P(\mathcal{X}^n) = 1$.
- Como $A_\epsilon^{(n)}$ é menor, teremos $P(A_\epsilon^{(n)}) \approx 1$.
- Existe um outro conjunto menor que contenha 'toda' a probabilidade? Todos os elementos em $A_\epsilon^{(n)}$ são essenciais? Se existir um conjunto menor, podemos criar um código utilizando este conjunto e assim obter uma melhor compressão.
- Veremos que $A_\epsilon^{(n)}$ é o menor conjunto com 'toda' a probabilidade.

Existem outro conjunto muito provável? II

- Vamos supor $B_\delta^{(n)}$ um conjunto qualquer com a propriedade

$$P(B_\delta^{(n)}) \geq 1 - \delta. \tag{258}$$

$B_\delta^{(n)}$ contém as sequências mais prováveis.

Teorema

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

$$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta' \text{ se } n \text{ for grande suficiente} \tag{259}$$

$$\Rightarrow |B_\delta^{(n)}| > 2^{n(H-\delta')} \approx 2^{nH} \tag{260}$$

- Assintoticamente $B_\delta^{(n)}$ não é menor do que $A_\epsilon^{(n)}$.

Existem outro conjunto muito provável? III

Definição

A notação $a_n \doteq b_n$ significa

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0, \tag{261}$$

então $a_n \doteq b_n$ implica que a_n e b_n são iguais até a primeira ordem do expoente. Ou seja, podemos dizer que para n grande, a_n e b_n possuem aproximadamente o mesmo comportamento. (Ver exemplos)

Podemos então reescrever o teorema anterior da seguinte forma:

Teorema

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

$$|B_{\delta_n}^{(n)}| \doteq |A_{\epsilon_n}^{(n)}| \doteq 2^{nH}. \tag{262}$$

Existem outro conjunto muito provável? IV

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 $\epsilon < \frac{1}{2}$. 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

$$P(A^c \cup B^c) \leq P(A^c) + P(B^c). \tag{263}$$

...

Existem outro conjunto muito provável? V

Demonstração.

continuação...

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

$$\begin{aligned} P(A \cap B) &= 1 - P(A^c \cup B^c) \\ &\geq 1 - P(A^c) - P(B^c) \\ &\geq 1 - \epsilon_1 - \epsilon_2. \end{aligned} \tag{264}$$

Podemos reescrever a desigualdade anterior como

$$\Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \geq 1 - \epsilon - \delta. \tag{265}$$

...

Existem outro conjunto muito provável? VI

Demonstração.

continuação...

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

$$\Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) = \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \tag{266}$$

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

...

Existem outro conjunto muito provável? VII

Demonstração.

Existem outro conjunto muito provável? VIII

continuação...

Desta forma teremos

$$\begin{aligned} 1-\epsilon-\delta &\leq \Pr(A_\epsilon^{(n)} \cap B_\delta^{(n)}) \\ &= \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \\ &\leq \sum_{x^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} \\ &= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} \\ &\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}, \end{aligned} \tag{267}$$

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

...

Existem outro conjunto muito provável? IX

Demonstração.

continuação...

Poderemos reescrever então da seguinte forma,

$$|B_{\delta}^{(n)}| > 2^{n(H-\epsilon)}, \tag{268}$$

onde $\epsilon > 0$.

□

Existem outro conjunto muito provável? X

Vejamos alguns exemplos para entender o significado de $a_n \stackrel{\circ}{=} b_n$.

Exemplo

Suponha que a sequência a_n seja definida como $a_n = e^{3n+1}$ e a b_n definida como $b_n = e^{3n}$. Teremos que

$$\log \frac{a_n}{b_n} = \log e, \tag{269}$$

ou seja, embora a_n e b_n sejam diferentes em todos os pontos, sendo a_n maior que b_n por um fator constante e, o logarítmo da razão entre ambas sequências não cresce muito rápido, o que pode ser verificado por

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{\log e}{n} = 0. \tag{270}$$

Existem outro conjunto muito provável? XI

Exemplo

Considere agora $a_n = e^{3n+\sqrt{n}}$ e $b_n = e^{3n}$. Então

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{1}{n}\sqrt{n} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} = 0. \tag{271}$$

Embora agora a razão entre a_n e b_n seja crescente com n, seu crescimento é lento.

Existem outro conjunto muito provável? XII

Exemplo

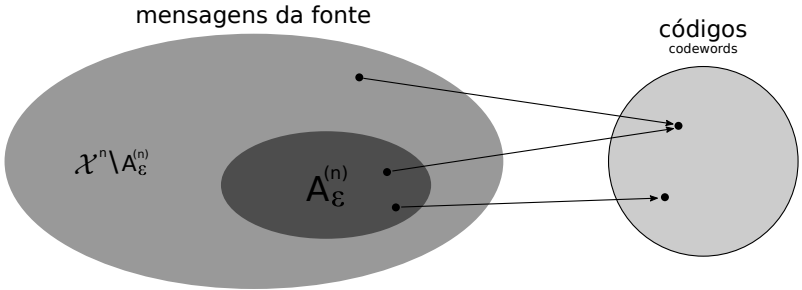
Seja $a_n = e^{4n}$ e $b_n = e^{3n}$. Teremos agora

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = \lim_{n\to\infty} \frac{1}{n} n = 1. \tag{272}$$

Neste caso, a razão entre a_n e b_n cresce muito rápido, o que foi constatado ao verificar que o limite da razão acima não convergiu para zero.

Estratégia de codificação com erros I

- Considere uma variação do código anterior que pode cometer erros.
- Sequências típicas utilizarão nH bits.
- Sequências atípicas serão mapeadas arbitrariamente em palavras curtas.



Estratégia de codificação com erros II

- Sabemos que $P(A_\epsilon^{(n)}) > 1 - \epsilon$.
- Um erro ocorre quando a sequência não é típica, logo a probabilidade de erro P_e é limitada por

$$P_e = P(A_\epsilon^{(n)c}) \leq \epsilon \tag{273}$$

- Conjunto típico: $\forall \epsilon > 0$, $\forall \delta > 0$, $\exists n_0$ tal que, para $n > n_0$,

$$p\left\{ | -\frac{1}{n} \log p(x_1, \ldots, x_n) - H| > \epsilon \right\} \leq \delta \tag{274}$$

  Podemos pensar como se fosse uma função $\delta(n, \epsilon)$ com $\lim_{n\to\infty} \delta(n, \epsilon) = 0$ para todo $\epsilon > 0$.
- Teremos então $P_e = P(A_\epsilon^{(n)c}) \leq \delta(n, \epsilon)$

$$P_e \to 0 \text{ à medida que } n \to \infty \tag{275}$$

Estratégia de codificação com erros III

- Independente de utilizarmos uma palavra longa e não tivermos erro, ou tivemos erro, o comprimento esperado é o mesmo e a probabilidade de erro vai pra zero se codificarmos o conjunto típico.

Codificando com menos do que H bits I

- O primeiro teorema de Shannon afirma que a codificação será sem erro se utilizarmos $n(H+\epsilon)$ bits por palavra código, para qualquer $\epsilon > 0$. O que ocorre se utilizarmos menos?
- Se utilizarmos $n(H-\alpha\epsilon)$ bits por palavra, com $\alpha > 1$, teremos no máximo $2^{n(H-\alpha\epsilon)}$ palavras.
- $2^{-n(H-\epsilon)}$ é o limite superior da probabilidade de sequências típicas.
- A probabilidade de sequências para as quais seremos capazes de fornecer uma palavra código não será maior do que o produto

$$2^{n(H-\alpha\epsilon)}2^{-n(H-\epsilon)} = 2^{-n\epsilon(\alpha-1)} \tag{276}$$

- Para qualquer $\alpha > 1$, esta probabilidade $\to 0$ quando $n \to \infty$.
- Problema: a probabilidade de sermos capazes de fornecer palavras código irá diminuir exponencialmente com n, pois a probabilidade da tipicidade diminui exponencialmente mais rápido do que o crescimento do número de códigos com n.
- O erro vai para 1 quando $n \to \infty$.

Codificando com menos do que H bits II

- Dadas n v.a.s com entropia H, podemos comprimi-las com mais do que nH bits com um risco mínimo de perder informação, quando $n \to \infty$. De maneira oposta, se as v.a.s são comprimidas a menos do que nH bits, então é virtualmente certo que haverá perda de informação e incorremos em erro.

Codificação do conjunto típico / Compressão (resumo) I

- Existe um código que pode alcançar uma taxa de compressão de $H(X) + \epsilon'$ bits por símbolo da fonte, para qualquer $\epsilon' > 0$, enquanto o comprimento do bloco n (número de símbolos da fonte que são codificados simultaneamente) for longo suficiente.
- Embora as sequências produzidas pelas fonte são consideradas i.i.d., é necessário codificá-las conjuntamente para atingir esta taxa.
- Fazemos isto utilizando $n(H + \epsilon) + 2$ bits para cada sequência típica e $n \log K + 2$ bits para cada sequência atípica. Este código é 1-1 e garantidamente sem erro.
- De maneira alternativa, podemos projetar para que ocorra um erro quando recebermos uma sequência atípica. O código possui uma probabilidade de erro P_e que não será nula, mas $\to 0$ exponencialmente rápido, quando $n \to \infty$, se $\epsilon' > 0$.
- Em qualquer caso, o comprimento esperado é

$$E[\frac{1}{n} l(X_{1:n})] \leq H(X) + \epsilon \tag{277}$$

Exercício 1 I

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

a) *(Desigualdade de Markov) Para qualquer v.a. não negativa X e qualquer $\delta > 0$, mostre que*

$$\Pr(X \geq \delta) \leq \frac{\mathrm{E}X}{\delta}. \tag{278}$$

Mostre uma v.a. para a qual teremos igualdade na equação acima.

...

Exercício 1 II

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

continuação...

***solução**: Se X possui distribuição $f(x)$, então*

$$\begin{aligned} \mathrm{E}X &= \int_0^\infty x f(x)\mathrm{d}x = \int_0^\delta x f(x)\mathrm{d}x + \int_\delta^\infty x f(x)\mathrm{d}x \\ &\geq \int_\delta^\infty x f(x)\mathrm{d}x \\ &\geq \int_\delta^\infty \delta f(x)\mathrm{d}x \\ &= \delta \Pr\left(X \geq \delta\right). \end{aligned} \tag{279}$$

...

Exercício 1 III

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

continuação...

Podemos assim concluir que

$$\Pr\left(X \geq \delta\right) \leq \frac{\mathrm{E}X}{\delta}. \tag{280}$$

...

Exercício 1 IV

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

continuação...

b) *(Desigualdade de Chebyshev) Seja Y uma v.a. com média μ e variância σ^2. Façamos $X = (Y - \mu)^2$. Mostre que para qualquer $\varepsilon > 0$,*

$$\Pr\left(|Y - \mu| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}. \tag{281}$$

...

Exercício 1 V

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

continuação...

***solução**:*

Utilizando a desigualdade de Markov com $X = (Y - \mu)^2$ *e* $\delta = \varepsilon^2$, *temos*

$$\Pr\left((Y-\mu)^2 \geq \varepsilon^2\right) \leq \frac{\mathrm{E}(Y-\mu)^2}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \tag{282}$$

Vamos utilizar também que $\Pr\left((Y-\mu)^2 > \varepsilon^2\right) \leq \Pr\left((Y-\mu)^2 \geq \varepsilon^2\right)$ *e* $\Pr\left((Y-\mu)^2 > \varepsilon^2\right) = \Pr\left(|Y-\mu| > \varepsilon\right)$. *Assim teremos*

$$\Pr\left(|Y-\mu| > \varepsilon\right) \leq \frac{\sigma^2}{\varepsilon^2}, \tag{283}$$

como queríamos demonstrar.

...

Exercício 1 VI

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

continuação...

c) *(Lei fraca dos grandes números) Seja $Z_1, Z_2, \ldots, Z_n$ uma sequência de v.a. i.i.d. com média μ e variância σ^2. Seja $\overline{Z_n} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, a média amostral. Mostre que*

$$\Pr\left(|\overline{Z_n} - \mu| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}. \tag{284}$$

Então teremos $\Pr\left(|\overline{Z_n} - \mu| > \varepsilon\right) \to 0$ quando $n \to \infty$, sendo esta a lei fraca dos grandes números.

...

Exercício 1 VII

Exercício (Desigualdade de Markov e Desigualdade de Chebyshev)

continuação...

***solução**:*

Vamos utilizar a desigualdade de Chebyshev com $Y = \overline{Z_n}$*, observado que* $\mathrm{E}\overline{Z_n} = \mu$ *e* $\mathrm{Var}(\overline{Z_n}) = \frac{\sigma^2}{n}$ *(i.e.* $\overline{Z_n}$ *é a soma de* n *v.a. i.i.d.* $\frac{Z_i}{n}$*, cada uma com variância* $\frac{\sigma^2}{n^2}$*). Teremos assim*

$$\Pr\left(|\overline{Z_n} - \mu| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}. \tag{285}$$

A desigualdade de Chebyshev é utilizada para provar a propriedade da equipartição assintótica (ver teorema 47).

Exercício 7 I

Exercício (PEA e codificação de fonte)

Uma fonte discreta sem memória emite uma sequência binária de dígitos independentes com probabilidade $p(1) = 0.005$ e $p(0) = 0.995$. Os dígitos são tomados em grupo de 100 e uma palavra binária é fornecida para cada sequência de 100 dígitos contendo três ou menos uns.

...

Exercício 7 II

Exercício (PEA e codificação de fonte)

continuação...

a) *Assumindo que todas as palavras código possuem a mesmo comprimento, encontre o comprimento mínimo necessário para fornecer código para todas as sequências com três ou menos uns (1s).*

solução

O número de sequências binárias de 100 bits com três ou menos uns é dado por

$$\binom{100}{0} + \binom{100}{1} + \binom{100}{2} + \binom{100}{3} = 1 + 100 + 4950 + 161700 = 166751 \tag{286}$$

...

Exercício 7 III

Exercício (PEA e codificação de fonte)

continuação...

Considerando que os códigos terão todos o mesmo comprimento, que deverá ser $\lceil \log 166751 \rceil = 18$. *Note que* $H(0.005) = 0.0454$, *logo para uma sequência de 100 símbolos teremos* 4.5 *bits de entropia, que é bem menor do que os 18 bits encontrados.*

...

Exercício 7 IV

Exercício (PEA e codificação de fonte)

continuação...

b) *Calcule a probabilidade de observar uma sequência da fonte para a qual nenhum código foi atribuído.*

solução

Devemos considerar aqui as sequências com mais de 3 uns. A probabilidade observarmos uma sequência para a qual não existe código associado é igual a um menos a probabilidade de observármos uma sequência para a qual existe código associado, ou seja,

$$1-\sum_{i=0}^{3}\binom{100}{i}(0.005)^i(0.995)^{100-i}=1-0.60577-0.30441-0.01243=0.00167. \quad (287)$$

...

Exercício 7 V

Exercício (PEA e codificação de fonte)

continuação...

c) *Use a desigualdade de Chebyshev para limitar a probabilidade de observamos um sequencia da fonte para a qual nenhum código foi associado. Compare este limite com o valor calculado no item anterior.*

solução

Para a v.a. S_n que representa a soma das v.a. i.i.d. $X_1, \ldots, X_n$, a desigualdade de Chebyschev afirma que

$$\Pr\left(|S_n - n\mu| \geq \epsilon\right) \leq \frac{n\sigma^2}{\epsilon^2}, \tag{288}$$

onde μ e σ^2 representam a média e variância de X_i (logo, $n\mu$ e $n\sigma^2$ são a média e variância de S_n). ...

Exercício 7 VI

Exercício (PEA e codificação de fonte)

continuação...

No problema em questão temos $n = 100$, $\mu = 0.005$ *e* $\sigma^2 = (0.005)(0.995)$. *Note que* $S_{100} \geq 4$ *se e somente se* $|S_{100} - 100(0.005)| \geq 3.5$. *Devemos então escolher* $\epsilon = 3.5$. *Então*

$$\Pr\left(S_{100} \geq 4\right) \leq \frac{100(0.005)(0.995)}{(3.5)^2} \approx 0.04061. \tag{289}$$

O limite encontrado é maior do que a real probabilidade 0.00167.

Exercício 8 I

Exercício (Comportamento limite do produto)

Seja a v.a.

$$X = \begin{cases} 1, & \text{com probabilidade } \frac{1}{2} \\ 2, & \text{com probabilidade } \frac{1}{4} \\ 3, & \text{com probabilidade } \frac{1}{4} \end{cases} . \tag{290}$$

Sejam $X_1, X_2, \ldots$ com a mesma distribuição. Encontre o comportamento limite do produto

$$P_n = (X_1 X_2 \ldots X_n)^{\frac{1}{n}}. \tag{291}$$

...

Exercício 8 II

Exercício (Comportamento limite do produto)

continuação...

solução

Tirando o logaritmo de P_n teremos

$$\log P_n = \frac{1}{n}\sum_{i=1}^{n} \log X_i \to \mathrm{E}\log X \tag{292}$$

com probabilidade 1, pela lei forte dos grandes números. Então $P_n \to 2^{\mathrm{E}\log X}$ com probabilidade 1. $\mathrm{E}\log X = \frac{1}{2}\log 1 + \frac{1}{4}\log 2 + \frac{1}{4}\log 3 = \frac{1}{4}\log 6$. *Logo,* $P_n \to 2^{\frac{1}{4}\log 6} = 1.565$.

Exercício 9 I

Exercício (Prop. da Eq. Ass.)

Seja $X_1, X_2, \ldots$ *v.a. independentes identicamente distribuídas com distribuição* $p(x)$, $x \in \{1, 2, \ldots, m\}$. *Então* $p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$. *Sabemos que* $-\frac{1}{n} \log p(X_1, X_2, \ldots, X_n) \to H(X)$ *em probabilidade. Seja* $q(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i)$, *onde* q *é outra função massa de probabilidade em* $\{1, 2, \ldots, m\}$.

...

Exercício 9 II

Exercício (Prop. da Eq. Ass.)

continuação...

a) *Avalie* $\lim -\frac{1}{n}\log q(X_1, X_2, \ldots, X_n)$, *onde* $X_1, X_2, \ldots$ *são i.i.d.* $\sim p(x)$.

solução

Como $X_1, X_2, \ldots, X_n$ *são i.i.d., então também serão* $q(X_1), q(X_2), \ldots, q(X_n)$. *Poderemos assim aplicar a lei forte dos grandes números*

...

Exercício 9 III

Exercício (Prop. da Eq. Ass.)

continuação...

$$\begin{aligned}
\lim -\frac{1}{n}\log q(X_1, X_2, \ldots, X_n) &= \lim -\frac{1}{n}\sum \log q(X_i) \\
&= -\mathrm{E}\log q(X) \\
&= -\sum p(x)\log q(x) \\
&= \sum p(x)\log\frac{p(x)}{q(x)} - \sum p(x)\log p(x) \\
&= D(p\,||\,q) + H(p). \qquad (293)
\end{aligned}$$

...

Exercício 9 IV

Exercício (Prop. da Eq. Ass.)

continuação...

b) *Agora avalia o limite da razão do logarítmo da verossimilhança* $\frac{1}{n}\log\frac{q(X_1,\ldots,X_n)}{p(X_1,\ldots,X_n)}$ *quando* $X_1, X_2, \ldots$ *são i.i.d.* $\sim p(x)$. *Então a chance de favorecer* q *é exponencialmente pequena quando* p *é verdadeiro.*

...

Exercício 9 V

Exercício (Prop. da Eq. Ass.)

continuação...

solução

Utilizando novamente a lei forte dos grandes números, teremos

$$\begin{aligned}
\lim -\frac{1}{n}\log\frac{q(X_1,X_2,\ldots,X_n)}{p(X_1,X_2,\ldots,X_n)} &= \lim -\frac{1}{n}\sum\log\frac{q(X_i)}{p(X_i)} \\
&= -\mathrm{E}\left(\log\frac{q(X)}{p(X)}\right) = -\sum p(x)\log\frac{q(X)}{p(X)} \\
&= \sum p(x)\log\frac{p(X)}{q(X)} \\
&= D(p\,||\,q) \qquad (294)
\end{aligned}$$

Exercício 11 I

Exercício (Prova do Teorema)

Seja $X_1, X_2, \ldots, X_n$ *i.i.d.* $\sim p(x)$. *Seja* $B_\delta^{(n)} \subset \mathcal{X}^n$ *tal que* $\Pr\left(B_\delta^{(n)}\right) > 1 - \delta$. *Fixe* $\epsilon < \frac{1}{2}$.

a) *Dados dois subconjuntos* A *e* B, *tais que,* $\Pr(A) > 1 - \epsilon_1$ *e* $\Pr(B) > 1 - \epsilon_2$, *mostre que* $\Pr(A \cap B) > 1 - \epsilon_1 - \epsilon_2$. *Então* $\Pr\left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) \geq 1 - \epsilon - \delta$.

...

Exercício 11 II

Exercício (Prova do Teorema)

continuação...

solução

Seja A^c o complemento de A. Então

$$\Pr\left(A^c \cup B^c\right) \leq P(A^c) + P(B^c). \tag{295}$$

Como $\Pr(A) \geq 1 - \epsilon_1$, $\Pr(A^c) \leq \epsilon_1$. *De forma similiar,* $\Pr(B^c) \leq \epsilon_2$. *Então*

$$\begin{aligned} \Pr(A \cap B) &= 1 - \Pr(A^c \cup B^c) \\ &\geq 1 - \Pr(A^c) - \Pr(B^c) \\ &\geq 1 - \epsilon_1 - \epsilon_2. \end{aligned} \tag{296}$$

...

Exercício 11 III

Exercício (Prova do Teorema)

continuação...

b) *Justifique os passos.*

solução

$$
\begin{aligned}
1-\epsilon-\delta &\leq \Pr\left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) \\
&\quad \textit{utilizando o item anterior do exercício} \\
&= \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} p(x^n) \\
&\quad \textit{definição de prob. de um conj.} \\
&\leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} \qquad (297) \\
&\quad \textit{limite da prob. dos elementos no conj. típico}
\end{aligned}
$$

...

Exercício 11 IV

Exercício (Prova do Teorema)

continuação...

$$
\begin{aligned}
1-\epsilon-\delta &\leq \sum_{A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} \\
&= |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)} \\
&\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)} \\
&\quad \textit{pois } \left(A_\epsilon^{(n)} \cap B_\delta^{(n)}\right) \subseteq B_\delta^{(n)}. \qquad (298)
\end{aligned}
$$

Exercício 3 - *Piece of Cake* I

Exercício (Piece of Cake)

Um bolo é partido em dois, conforme as proporções abaixo.

$$P = \begin{cases} \left(\frac{2}{3}, \frac{1}{3}\right), & \text{com prob.} \frac{3}{4}, \\ \left(\frac{2}{5}, \frac{3}{5}\right), & \text{com prob.} \frac{1}{4}. \end{cases} \tag{299}$$

A maior metade é escolhida e a menor descartada. A metade é subsequentemente redividida seguindo as mesmas regras. Exemplo: o primeiro corte pode resultar em uma fatia de tamanho $\frac{3}{5}$*, um novo corte poderia resultar em um pedaço de tamanho* $\left(\frac{3}{5}\right)\left(\frac{2}{3}\right)$*, e assim por diante, Qual é o tamanho, até a primeira ordem do expoente, do pedaço de bolo remanescente após* n *cortes sucessivos?*

...

Exercício 3 - *Piece of Cake* II

Exercício (Piece of Cake)

continuação...

solução *Vamos chamar de C_i a fração do pedaço de bolo decorrente do i-ésimo corte, e chamaremos de T_n a fração remanescente do bolo após n cortes. Teremos então $T_n = C_1 C_2 \dots C_n = \prod_{i=1}^{n} C_i$.*

Temos que

$$\begin{aligned} \lim \frac{1}{n} \log T_n &= \lim \frac{1}{n} \sum_{i=1}^{n} \log C_i \\ &\quad \text{pela lei forte dos grandes números e como } C_i \text{ são i.i.d. teremos} \\ &= \mathrm{E}\left[\log C_1\right] \\ &= \frac{3}{4} \log \frac{2}{3} + \frac{1}{4} \log \frac{3}{5} = -0.62296. \end{aligned} \tag{300}$$

Teremos então $T_n \to 2^{n \mathrm{E}[\log C_1]}$.

Exercício 10 - *Random box size I*

Exercício (Tamanho da caixa aleatória)

Uma caixa retangular aleatória n dimensional com lados $X_1, \ldots, X_n$ é construída. O volume desta caixa é $V_n = \prod_{i=1}^{n} X_i$. O comprimento de aresta de um cubo n dimensional com mesmo volume que a caixa aleatória é dado por $l = V_n^{1/n}$. Seja $X_1, X_2, \ldots$ v.a. i.i.d. com distribuição uniforme em $[0, 1]$. Encontre o limite $\lim_{n\to\infty} V_n^{1/n}$ e compare com $(\mathrm{E}V_n)^{1/n}$. Veremos que o valor esperado do comprimento da aresta não captura a idéia do volume da caixa. A média geométrica, ao invés da média aritmética, caracteriza o comportamento de produtos.

...

Exercício 10 - *Random box size II*

Exercício (Tamanho da caixa aleatória)

continuação...

solução *O volume da caixa* $V_n = \prod_{i=1}^{n} X_i$ *é uma v.a., já que* X_i *são v.a.* $\sim u(0,1)$. V_n *tende a* 0 *quando* $n \to \infty$.
Utilizando a lei forte dos grandes números e o fato de X_i *serem i.i.d., temos*

$$\begin{aligned} \ln V_n^{1/n} &= \frac{1}{n} \ln V_n \\ &= \frac{1}{n} \sum \ln X_i \to \mathrm{E}\left[\ln(X)\right] \end{aligned} \tag{301}$$

Temos ainda que

$$\mathrm{E}\left[\ln(X)\right] = \int_0^1 \ln x \mathrm{d}x = (x \ln x - x)|_{p=0}^{p=1} = -1. \tag{302}$$

...

Exercício 10 - *Random box size III*

Exercício (Tamanho da caixa aleatória)

continuação...

$$\begin{aligned} \lim_{n\to\infty} V_n^{1/n} &= \lim_{n\to\infty} e^{\frac{1}{n}\ln V_n} \\ &\quad \textit{como } e^x \textit{ é uma função contínua} \\ &= e^{\lim_{n\to\infty}\frac{1}{n}\ln V_n} = e^{-1} < \frac{1}{2}. \end{aligned} \tag{303}$$

$\frac{1}{2}$ *é a média aritmética da v.a. e* $\frac{1}{e}$ *é a média geométrica. O volume esperado da caixa é* $\mathrm{E}(V_n) = \prod_{i=1}^{n} \mathrm{E}X_i = \left(\frac{1}{2}\right)^n$.

Exercício 13 - Cálculo do conjunto típico I

Exercício (Cálculo do conjunto típico)

Considere uma sequência i.i.d. de v.a. binárias $X_1, \ldots, X_n$ *com distribuição Bern*(p) *onde* $p = 0.6$, *ou seja,* $\Pr(X_i = 1) = 0.6$ *(logo,* $\Pr(X_i = 0) = 0.4$*).*

a) *Calcule* $H(X)$

solução

$$H(X) = H(p) = -0.6 \log 0.6 - 0.4 \log 0.4 = 0.97095 \textit{bits}. \tag{304}$$

...

Exercício 13 - Cálculo do conjunto típico II

Exercício (Cálculo do conjunto típico)

continuação...

b) *Considere $n = 25$ e $\epsilon = 0.1$. Quais sequências pertencem ao conjunto típico $A_\epsilon^{(n)}$? Qual é a probabilidade do conjunto típico? Quantos elementos pertencem ao conjunto típico?*

solução *Pela definição, o conjunto típico é o conjunto das sequências em que $-\frac{1}{n}\log p(x^n)$ está no intervalo $(H(X) - \epsilon, H(X) + \epsilon)$, isto é, $(0.87095, 1.07095)$. A probabilidade de uma sequência depende do seu histograma empírico, neste caso, do número de ocorrências de 1 e 0, sendo dada por*

$$p(x^n) = p^k(1-p)^{n-k}, \tag{305}$$

onde k é o número de ocorrências de 1. Vamos resolver computacionalmente: ...

Exercício 13 - Cálculo do conjunto típico III

Exercício (Cálculo do conjunto típico)

continuação...

```
» p=0.6;
» H=-p*log2(p)-(1-p)*log2(1-p);
» eps=0.1; n=25;
» for k=0:n,
pn=p^k*(1-p)^(n-k); lpn=-(1/n)*log2(pn);
if lpn>H-eps && lpn<H+eps, disp(k); endif;
endfor;
11 12 13 14 15 16 17 18 19
```

...

Exercício 13 - Cálculo do conjunto típico IV

Exercício (Cálculo do conjunto típico)

continuação...

Agora que determinamos quais sequências pertencem ao conjunto típico, podemos calcular a probabilidade deste conjunto.

$$\Pr(A_\epsilon^{(n)}) = \sum_{k=11}^{19} \binom{n}{k} p^k (1-p)^{n-k}. \tag{306}$$

...

Exercício 13 - Cálculo do conjunto típico V

Exercício (Cálculo do conjunto típico)

continuação...

Iremos novamente resolver computacionalmente.

```
» PA=0;
» for k=0:n,
pn=p^k*(1-p)^(n-k); lpn=-(1/n)*log2(pn);
if lpn>H-eps && lpn<H+eps,
PA+=nchoosek(n,k)*pn;
endif;
endfor; » PA = 0.93625
```

Note que o valor de $\Pr(A_\epsilon^{(n)})$ *encontrado é maior do que* $1-\epsilon$*, ou seja,* n *dado é grande suficiente para garantir que* $\Pr(A_\epsilon^{(n)}) > 1-\epsilon$.

...

Exercício 13 - Cálculo do conjunto típico VI

Exercício (Cálculo do conjunto típico)

continuação...

O número de sequências no conjunto típico, ou seja, o tamanho do conjunto é dado por

$$|A_\epsilon^{(n)}| = \sum_{k=11}^{19} \binom{n}{k} \tag{307}$$

```
» TA=0;
for k=0:n,
pn=p^k*(1-p)^(n-k); lpn=-(1/n)*log2(pn);
if lpn>H-eps && lpn<H+eps, TA+=nchoosek(n,k); endif;
endfor; TA
TA = 26366510
```

...

Exercício 13 - Cálculo do conjunto típico VII

Exercício (Cálculo do conjunto típico)

continuação...

Podemos ainda calcular os limites do tamanho deste conjunto:

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}. \tag{308}$$

```
» ll=(1-eps)*2^(n*(H-eps))
ll = 3226999.15174967
» LL=2^(n*(H+eps))
LL = 114737747.617766
» TA > ll && TA < LL
ans = 1
```

Método de Tipo I

- Uma refinamento da abordagem das sequências típicas.
- Dado $X_1, X_2, \ldots, X_n$ i.i.d. $\sim p(x)$, iremos particionar as D^n sequências em classes de acordo com a sua distribuição empírica (histograma), i.e., tipo da sequência. Obs: $|\mathcal{X}| = D$.
- Uma classe de tipo é uma classe (ou conjunto) de sequências de comprimento n que possuem o mesmo histograma empírico.
- O número de tipos cresce sub-exponencialmente com n.
- Sequências do mesmo tipo são equiprováveis.
- O número de sequencias de uma certa classe de tipo cresce exponencialmente.
- A intersecção entre eventos de erro e eventos de classe de tipo permite encontrar bons limites para o erro.
- Iremos obter o teorema de Shannon (e a proposição inversa) de maneira formal mas intuitiva.

Definição de Tipo I

- 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.

$$P_{x_{1:n}} \triangleq \left(\frac{n(a_1|x_{1:n})}{n}, \frac{n(a_2|x_{1:n})}{n}, \ldots, \frac{n(a_D|x_{1:n})}{n} \right) \tag{309}$$

  onde $n(a_i|x_{1:n})$ representa o número de ocorrências do símbolo a_i na amostra $x_{1:n}$.
- $P_{x_{1:n}}$ é uma função massa probabilidade.
- $P_{x_{1:n}}$ é o histograma, ou tipo, da amostra.
- $P_{x_{1:n}}(a) = \frac{n(a|x_{1:n})}{n}$ para $a \in \mathcal{X}$.

Conjunto de Tipos I

- Vamos definir $\mathcal{P}_n$ como o conjunto de todas possíveis tipos com denominador n.
- $\mathcal{P}_n \equiv \mathcal{P}_n(\mathcal{X}) \equiv \mathcal{P}_n(|\mathcal{X}|)$ é o conjunto de tipos que podem ocorrer em sequências de comprimento n utilizando símbolos do alfabeto $\mathcal{X}$.
- Exemplo. $\mathcal{X} = \{0, 1\}$, então

$$\mathcal{P}_n(\mathcal{X}) = \left\{ \left(\frac{0}{n}, \frac{n}{n}\right), \left(\frac{1}{n}, \frac{n-1}{n}\right), \ldots, \left(\frac{n}{n}, \frac{0}{n}\right) \right\} \tag{310}$$

  Neste caso existe no total $n + 1$ tipos (histogramas).
- Observação: note que $\mathcal{P}_n$ é um conjunto de listas ordenadas. Usualmente utilizamos $\{\cdot\}$ para designar conjuntos e $(\cdot)$ para designar listas ordenadas.

Classe de Tipo I

- Para um dado tipo $P \in \mathcal{P}_n$, o conjunto de sequências de comprimento n do tipo P constitui o que chamamos de classe de tipo de P.
- Será designado por $T(P)$.

$$T(P) \triangleq \{x_{1:n} \in \mathcal{X}^n : P_{x_{1:n}} = P\} \tag{311}$$

que é o conjunto de todas as sequências de comprimento n com um determinado tipo (histograma) P.

Sumário I

Dada uma sequencia de comprimento n, temos:

1) o tipo (ou histograma da amostra $x_{1:n}$

$$P_{x_{1:n}} \triangleq \left(\frac{n(a_1|x_{1:n})}{n}, \frac{n(a_2|x_{1:n})}{n}, \ldots, \frac{n(a_D|x_{1:n})}{n} \right) \tag{312}$$

2) o conjunto de todos os tipo (ou histogramas) $\mathcal{P}_n$
3) um tipo em particular $P \in \mathcal{P}_n$
4) classe de tipo: dado um tipo P, o conjunto de todas sequencias deste tipo

$$T(P) \triangleq \{x_{1:n} \in \mathcal{X}^n : P_{x_{1:n}} = P\} \tag{313}$$

Exemplo I

- Seja $\mathcal{X} = \{1, 2, 3\}$ e $x_{1:5} = [1, 1, 3, 2, 1]$.
- Então

$$P_{x_{1:5}} = \left(\frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right) \tag{314}$$

- $T(P_{x_{1:5}})$ é o conjunto de sequencias de comprimento 5 que possua três 1s, um 2, e um 3, i.e.,

$$T(P_{x_{1:5}}) = \{[1, 1, 1, 2, 3], [1, 1, 1, 3, 2], \ldots, [3, 2, 1, 1, 1]\} \tag{315}$$

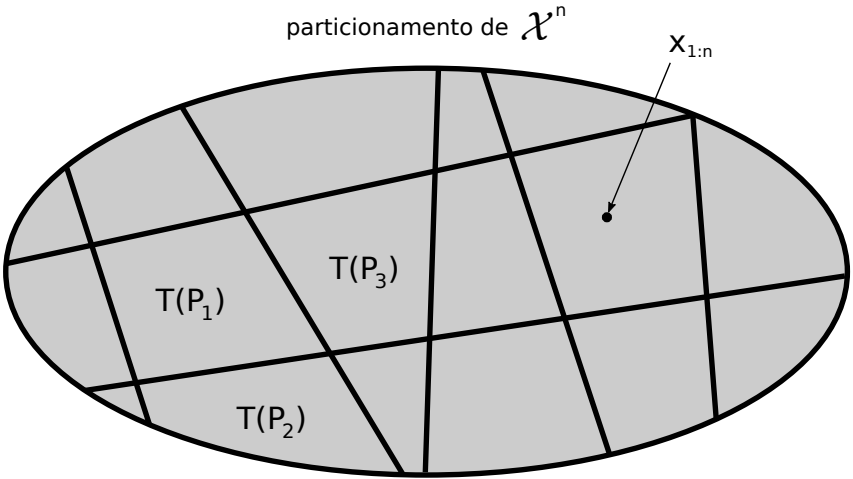
- Quantos tipos existem? Qual é o valor de $|\mathcal{P}_n|$? Neste exemplo, $|\mathcal{P}_n| = 21$.
- De forma geral, temos

$$|\mathcal{P}_n| = \binom{n + |\mathcal{X}| - 1}{|\mathcal{X}| - 1} \tag{316}$$

Divisão do Conjunto de Sequências em Classes de Tipo I

- $\mathcal{X}^n$ é o conjunto de todas as sequências de comprimento n.
- particionamento de $\mathcal{X}^n$.
- $T(P_i) \cap T(P_j) = \emptyset$, $\forall i, j \in \{1, \ldots, |\mathcal{P}_n|\}$ e $i \neq j$.
- $\mathcal{P}_n = \{P_1, P_2, \ldots, P_{|\mathcal{P}_n|}\}$ é o conjunto de todos os tipos.
- $\bigcup_{P \in \mathcal{P}_n} T(P) = \mathcal{X}^n$.

Divisão do Conjunto de Sequências em Classes de Tipo II



Limite no Número de Tipos I

Teorema (Limite no número de tipos)

O número de tipos para sequências de comprimento n em um alfabeto $\mathcal{X}$ é limitado por

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}. \tag{317}$$

Demonstração.

Limite no Número de Tipos II

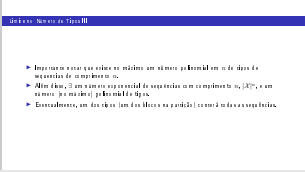
- Note que o numerador de cada entrada em um tipo pode assumir $(n+1)$ valores distintos (de 0 a n).
- Existem $|\mathcal{X}|$ entradas em um tipo, e portanto a mesma quantidade de numeradores.
- Os valores dos numeradores interagem entre si (a soma de todos deve ser igual a n), mas podemos achar um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

$$|\mathcal{P}_n| \leq \underbrace{(n+1) \times (n+1) \times \ldots \times (n+1)}_{|\mathcal{X}| \text{ vezes}} = (n+1)^{|\mathcal{X}|}. \tag{318}$$

□

Limite no Número de Tipos III

- Importante notar que existe no máximo um número polinomial em n de tipos de sequencias de comprimento n.
- Além disso, $\exists$ um número exponencial de sequências com comprimento n, $|\mathcal{X}|^n$, e um número (no máximo) polinomial de tipos.
- Eventualmente, um dos tipos (um dos blocos na partição) conterá todas as sequências.



Note que calcular Equação 316 é muito mais complicado do que calcular 317, e como visto o limite será suficiente para o que queremos mostrar.

Probabilidade Depende do Tipo I

Teorema (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,

$$Q^n(x_{1:n}) = 2^{-n[H(P_{x_{1:n}}) + D(P_{x_{1:n}} || Q)]} \tag{319}$$

- Probabilidade não depende da sequencia, dado o tipo.
- Estatística Suficiente.
- Todas as sequencias do mesmo tipo possuem a mesma probabilidade.

Probabilidade Depende do Tipo II

Demonstração.

Probabilidade Depende do Tipo III

$$
\begin{aligned}
Q^n(x_{1:n}) &= \prod_{i=1}^{n} Q(x_i) = \prod_{a\in\mathcal{X}} Q(a)^{n(a|x_{1:n})} \\
&= \prod_{a\in\mathcal{X}} Q(a)^{nP_{x_{1:n}}(a)} = \prod_{a\in\mathcal{X}} 2^{\left\{nP_{x_{1:n}}(a)\log Q(a)\right\}} \\
&= \prod_{a\in\mathcal{X}} 2^{n\left\{P_{x_{1:n}}(a)\log Q(a)\underbrace{-P_{x_{1:n}}(a)\log P_{x_{1:n}}(a)+P_{x_{1:n}}(a)\log P_{x_{1:n}}(a)}_{=0}\right\}} \\
&= 2^{n\sum_{a\in\mathcal{X}}\left(-P_{x_{1:n}}(a)\log\frac{P_{x_{1:n}}(a)}{Q(a)}+P_{x_{1:n}}(a)\log P_{x_{1:n}}(a)\right)} \\
&= 2^{-n\left(D(P_{x_{1:n}}||Q)+H(P_{x_{1:n}})\right)} \qquad (320)
\end{aligned}
$$

□

Probabilidade Depende do Tipo IV

- (corolário) Se Q é uma distribuição racional (i.e., um tipo possível) e se $x_{1:n} \in T(Q)$, então
$$Q^n(x_{1:n}) = 2^{-nH(Q)}. \tag{321}$$
- O que ocorre se Q for irracional? Podemos fazer $D(P_{x_{1:n}} \,||\, Q)$ tão pequeno quando desejável, fazendo n grande suficiente.

Classe de Tipo com maior probabilidade I

- Qual classe de tipo possui maior probabilidade quando a distribuição geradora das sequências é $P \in \mathcal{P}_n$?
- Considerando a Prop. da Eq. Ass., as sequências típicas são aquelas mais próximas da real distribuição, e elas possuem 'toda' probabilidade.
- Vamos supor então que $T(P)$ possui a maior probabilidade sob a distribuição P.

Lema

Para $P \in \mathcal{P}_n$, teremos que $T(P)$ possui a maior probabilidade. Isto é

$$P^n(T(P)) \geq P^n(T(\hat{P})),\ \forall \hat{P} \in \mathcal{P}_n. \tag{322}$$

Nota: 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 $m = n$, $\frac{m!}{n!} = 1 = n^0$.

Classe de Tipo com maior probabilidade II

Demonstração.

Classe de Tipo com maior probabilidade III

$$
\begin{aligned}
\frac{P^n(T(P))}{P^n(T(\hat{P}))} &= \frac{|T(P)| \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{|T(\hat{P})| \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \\
&= \frac{\binom{n}{nP(a_1)\ nP(a_2)\ \ldots\ nP(a_D)} \prod_{a\in\mathcal{X}} P(a)^{nP(a)}}{\binom{n}{n\hat{P}(a_1)\ n\hat{P}(a_2)\ \ldots\ n\hat{P}(a_D)} \prod_{a\in\mathcal{X}} P(a)^{n\hat{P}(a)}} \\
&= \prod_{a\in\mathcal{X}} \frac{[n\hat{P}(a)]!}{[nP(a)]!} P(a)^{n(P(a)-\hat{P}(a))} \\
&\geq \prod_{a\in\mathcal{X}} (nP(a))^{n(\hat{P}(a)-P(a))} P(a)^{n(P(a)-\hat{P}(a))} \\
&= \prod_{a\in\mathcal{X}} n^{n(\hat{P}(a)-P(a))} \qquad (323)
\end{aligned}
$$

...

Classe de Tipo com maior probabilidade IV

Demonstração.

continuação...

$$\begin{aligned}\frac{P^n(T(P))}{P^n(T(\hat{P}))} &= \dots \\ &\geq \prod_{a\in\mathcal{X}} n^{n(\hat{P}(a)-P(a))} \\ &= n^{n\left[\sum_{a\in\mathcal{X}}\hat{P}(a)-\sum_{a\in\mathcal{X}}P(a)\right]} \\ &= n^{n(1-1)} = 1 \end{aligned} \tag{324}$$

logo, $P^n(T(P)) \geq P^n(T(\hat{P}))$. □

Tamanho da Classe de Tipo I

Podemos expressar o tamanho de uma classe de tipo utilizando os coeficientes multinomiais, i.e., o número de maneiras de escolher símbolos distintos do alfabeto para cada elemento da sequencia $x_{1:n}$.
Para $P \in \mathcal{P}_n$, temos

$$|T(P)| = \binom{n}{nP(a_1)\ nP(a_2)\ \ldots\ nP(a_n)} \tag{325}$$

Entretanto, isto é difícil calcular. Queremos encontrar limites que sejam mais facilmente manipulados matematicamente.

Tamanho da Classe de Tipo II

Teorema (Limites no tamanho da Classe de Tipo)

Dado um tipo $P \in \mathcal{P}_n$, temos

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \tag{326}$$

limite superior.

$$\begin{aligned} 1 &\geq P^n(T(P)) = \sum_{x_{1:n} \in T(P)} P^n(x_{1:n}) = \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \\ &= |T(P)| 2^{-nH(P)} \end{aligned} \tag{327}$$

□

Tamanho da Classe de Tipo III

limite inferior.

$$
\begin{aligned}
1 &= \sum_{Q\in\mathcal{P}_n} P^n(T(Q)) \leq \sum_{Q\in\mathcal{P}_n} \max_{R\in\mathcal{P}_n} P^n(T(R)) \\
&\quad \text{fazendo } P = \operatorname*{argmax}_{R\in\mathcal{P}_n} P^n(T(R)) \text{ teremos} \\
&= \sum_{Q\in\mathcal{P}_n} P^n(T(P)) \leq (n+1)^{|\mathcal{X}|} P^n(T(P)) \\
&= (n+1)^{|\mathcal{X}|} \sum_{x_{1:n}\in T(P)} P^n(x_{1:n}) \\
&= (n+1)^{|\mathcal{X}|} \sum_{x_{1:n}\in T(P)} 2^{-nH(P)} \qquad (328)
\end{aligned}
$$

...

Tamanho da Classe de Tipo IV

limite inferior.

continuação...

$$\begin{aligned} 1 &= \dots \\ &\leq (n+1)^{|\mathcal{X}|} \sum_{x_{1:n} \in T(P)} 2^{-nH(P)} \\ &= (n+1)^{|\mathcal{X}|} |T(P)| 2^{-nH(P)} \end{aligned} \tag{329}$$

fornecendo assim o resultado

$$|T(P)| \geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \tag{330}$$

□

Limites Combinatórios I

- Para o caso binário, $\mathcal{X} = \{0, 1\}$, temos os seguintes limites

$$\frac{1}{(n+1)^2} 2^{nH(\frac{k}{n})} \leq \binom{n}{k} \leq 2^{nH(\frac{k}{n})} \tag{331}$$

- O limite inferior pode ser ainda mais restrito neste caso

$$\frac{1}{(n+1)} 2^{nH(\frac{k}{n})} \leq \binom{n}{k} \leq 2^{nH(\frac{k}{n})} \tag{332}$$

Probabilidade da classe de tipo I

Teorema

Para qualquer $P \in \mathcal{P}_n$ e qualquer distribuição Q, 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

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(P||Q)} \leq Q^n(T(P)) \leq 2^{-nD(P||Q)} \tag{333}$$

obs.: qualquer tipo menos próximo do tipo mais próximo de Q irá ter probabilidade exponencialmente decrescente com n, decrescendo mais rapidamente que o tipo mais provável.

Probabilidade da classe de tipo II

Demonstração.

$$\begin{aligned} Q^n(T(P)) &= \sum_{x_{1:n}\in T(P)} Q^n(x_{1:n}) = \sum_{x_{1:n}\in T(P)} 2^{-n(D(P||Q)+H(P))} \\ &= |T(P)|2^{-n(D(P||Q)+H(P))} \end{aligned} \tag{334}$$

para completar a demonstração, devemos utilizar

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)} \tag{335}$$

□

Probabilidade da classe de tipo III

- Quais tipos terão maior probabilidade?
- Claramente aqueles mais próximos da real distribuição.
- A propriedade $Q^n(T(P)) \doteq 2^{-nD(P||Q)}$ diz que aqueles mais distantes terão probabilidade exponencialmente menor do que os demais, quando $n \to \infty$.

Sumário I

- Número de tipos de sequências com comprimento n

$$|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|} \tag{336}$$

- A probabilidade da sequência ($p(x_{1:n})$) depende apenas do tipo

$$Q^n(x_{1:n}) = 2^{-n[H(P_{x_{1:n}})+D(P_{x_{1:n}}||Q)]} \tag{337}$$

- Tamanho da Classe de Tipo

$$|T(P)| \doteq 2^{nH(P)} \tag{338}$$

- Probabilidade da classe de tipo

$$Q^n(T(P)) \doteq 2^{-nD(P||Q)} \tag{339}$$

Conjunto Típico I

- Os tipos P mais próximos da real distribuição Q terão maior probabilidade.
- Aqueles mais distantes de Q terão probabilidade exponencialmente menor do que os demais.

Definição (conjunto típico de sequências)

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

$$T_Q^\epsilon = \{x_{1:n} : D(P_{x_{1:n}} \,||\, Q) \leq \epsilon\} \tag{340}$$

Conjunto Típico II

Teorema (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

$$Q(\overline{T}_Q^\epsilon) = Q(\{x_{1:n} : D(P_{x_{1:n}} \,||\, Q) > \epsilon\}) \leq 2^{-n(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n})} \tag{341}$$

desta forma

$$D(P_{x_{1:n}} \,||\, Q) \xrightarrow{p} 0 \text{ quando } n \to \infty \tag{342}$$

(converge em probabilidade para zero quando n é grande suficiente)

- 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.

Conjunto Típico III

Demonstração.

$$\begin{aligned} 1 - Q^n(T_Q^\epsilon) &= Q(\overline{T}_Q^\epsilon) = \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} Q^n(T(P)) \\ &\leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-nD(P||Q)} \\ &\leq \sum_{P \in \mathcal{P}_n : D(P||Q) > \epsilon} 2^{-n\epsilon} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\epsilon} = 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n}\right)} \end{aligned} \tag{343}$$

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

Conjunto Típico IV

- Com qual frequência um evento atípico ocorre?
- Como $p(\overline{T}_Q^\epsilon) \leq 2^{-n\left(\epsilon - |\mathcal{X}| \frac{\log(n+1)}{n}\right)}$, uma sequencia em n exponencial decrescente. Esta será, desta forma, somável.

$$\infty > \sum_{n=1}^{\infty} p(D(P_{x_{1:n}} \,||\, Q) > \epsilon) = E_Q \left[\sum_{n=1}^{\infty} \mathbf{1}_{\{D(P_{X_{1:n}}||Q) > \epsilon\}} \right] \tag{344}$$

- O número esperado de vezes que o evento $D(P_{X_{1:n}} \,||\, Q) > \epsilon$ ocorre é finito, dentro de um conjunto infinito de possíveis ocorrências.

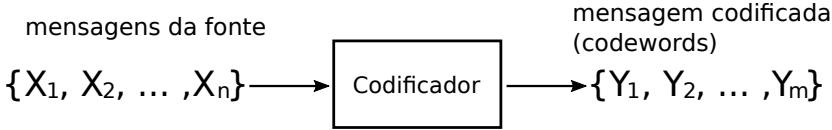
Codificação Universal de Fonte I

- Dizemos que uma codificação de fonte é universal quando ela não depende de $p(x)$.
- Codificação não universal: Se conhecemos $p(x)$, podemos propor um código para comprimir a fonte caracterizada por $p(x)$. Exemplo: Código de Huffman.
- É possível criar um código universal (não dependente de $p(x)$) que atinja o limite da entropia? taxa $R > H(Q)$ (em bits por símbolo)
- O que ocorre se $R < H(Q)$?
- Ideia similar àquela do conjunto típico $A_\epsilon^{(n)}$. Vamos codificar apenas aquilo que de fato ocorre. Precisaremos de no máximo $|A_\epsilon^{(n)}|$ palavras que podem ser indexadas com nH bits.
- Vamos formalizar o teorema de Shannon utilizando o método de tipo.

Codificação Universal de Fonte II

- 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.
- Existe um número exponencial (em n) de sequências e um número polinomial (em n) de tipos.
- Eventualmente, um tipo terá 'toda' a probabilidade.

Codificação Universal de Fonte III



$X_i \in \{a_1, a_2, \ldots, a_K\}$

K^n mensagens possíveis

n letras ou símbolos do alfabeto

$Y_i \in \{0,1\}$

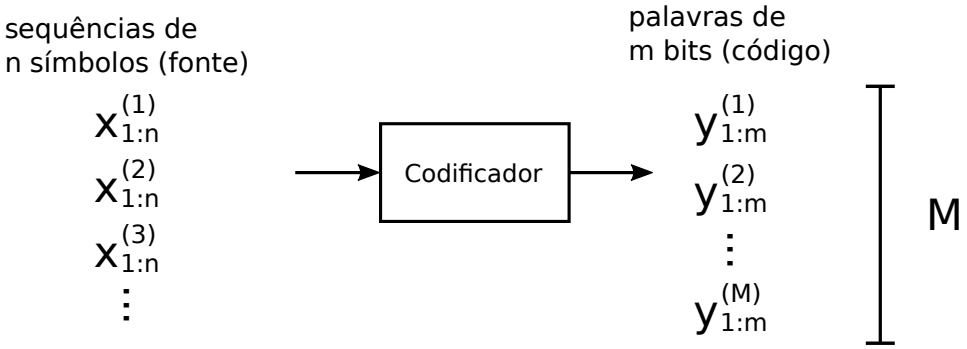
2^m mensagens possíveis

m total de bits

Códigos (M, n) I

- código de blocos com taxa fixa R
- Existem M palavras. M é o número de possíveis mensagens.
- n símbolos são codificados conjuntamente a cada instante.
- O codificador faz o mapeamento de *strings* de tamanho n produzidas pela fonte em *strings* de m bits.

Códigos (M, n) II



- A taxa R depende de M e n.

$$R = \frac{\log M}{n} = \frac{\log(\text{n. de palavras})}{\text{n. de símbols}} \tag{345}$$

Códigos (M, n) III

Definição (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:

$$\text{codificador: } f_n : \mathcal{X}^n \to \{1, 2, \ldots, 2^{nR}\} \tag{346}$$

$$\text{decodificador: } \phi_n : \{1, 2, \ldots, 2^{nR}\} \to \mathcal{X}^n \tag{347}$$

e a probabilidade de erro

$$P_e^{(n)} = Q^n(\{x_{1:n} : \phi(f_n(x_{1:n})) \neq x_{1:n}\}) \tag{348}$$

Códigos (M, n) IV

Definição (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

$$P_e^{(n)} \to 0 \text{ quando } n \to \infty \text{ sempre que } H(Q) < R \tag{349}$$

- 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.

Simplex Probabilístico I

Definição (Simplex Probabilístico)

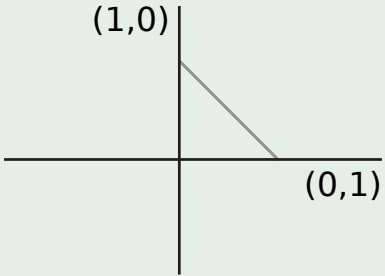
O Simplex Probabilístico em $\mathbb{R}^m$ é o conjunto de pontos $x_{1:m} = (x_1, x_2, \ldots, x_m) \in \mathbb{R}^m$ tal que $x_i \geq 0$, $\sum_{i=1}^{m} x_i = 1$.

Simplex Probabilístico II

Exemplo ($m = 2$)

O Simplex probabilístico será o conjunto de pontos

$$\{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 = 1\} \tag{350}$$

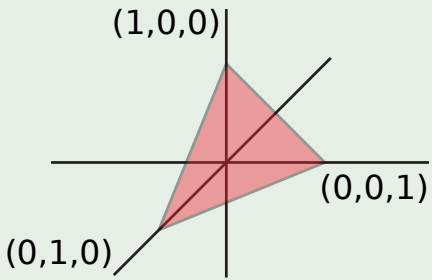


Simplex Probabilístico III

Exemplo ($m = 3$)

O Simplex probabilístico será o conjunto de pontos

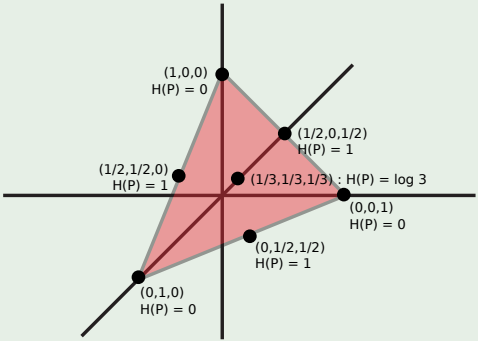
$$\{(x_1, x_2, x_3) : x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_1 + x_2 + x_3 = 1\} \tag{351}$$



Simplex Probabilístico IV

Os tipos em para $|\mathcal{X}| = m$ podem ser representados em um Simplex Probabilístico em $\mathbb{R}^m$.

Exemplo ($|\mathcal{X}| = 3$)



Teorema da Codificação de Shannon I

Teorema (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 $H(Q) < R$.

Teorema da Codificação de Shannon II

Demonstração.

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

$$R_n \triangleq R - |\mathcal{X}|\frac{\log(n+1)}{n} < R \tag{352}$$

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

$$\begin{aligned} A_n &\triangleq \{x_{1:n} \in \mathcal{X}^n : H(P_{x_{1:n}}) \leq R_n\} \\ &= \left\{ \bigcup_{P \in \mathcal{P}_n} T(P) : H(P) \leq R_n \right\} \end{aligned} \tag{353}$$

...

Teorema da Codificação de Shannon III

Demonstração.

continuação...

- Temos então

$$\begin{aligned} |A_n| &= \sum_{P\in\mathcal{P}_n:H(P)\le R_n} |T(P)| \le \sum_{P\in\mathcal{P}_n:H(P)\le R_n} 2^{nH(P)} \\ &\le \sum_{P\in\mathcal{P}_n:H(P)\le R_n} 2^{nR_n} \le (n+1)^{|\mathcal{X}|} 2^{nR_n} \\ &= 2^{n\left(R_n+|\mathcal{X}|\frac{\log(n+1)}{n}\right)} = 2^{nR}. \end{aligned} \tag{354}$$

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

...

Teorema da Codificação de Shannon IV

Demonstração.

continuação...

O codificador será dada por

$$f_n(x_{1:n}) = \begin{cases} \text{índice de } x_{1:n} \text{ em } A_n, & \text{se } x_{1:n} \in A_n \\ 0, & \text{caso contrário .} \end{cases} \tag{355}$$

- 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 $x_{1:n} \notin A_n$.

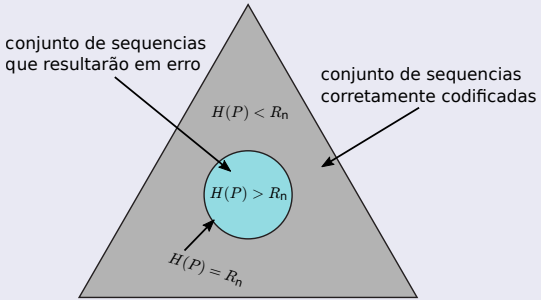
...

Teorema da Codificação de Shannon V

Demonstração.

continuação...

Os tipos podem ser representados por pontos em um Simplex Probabilístico.



...

Teorema da Codificação de Shannon VI

Demonstração.

continuação...

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

$$\begin{aligned} P_e^{(n)} 1 - Q^n(A_n) &= Q^n(A_n^c) = \sum_{P:H(P)>R_n} Q^n(T(P)) \\ &\leq \sum_{P:H(P)>R_n} \max_{P:H(P)>R_n} Q^n(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{P:H(P)>R_n} Q^n(T(P)) \\ &\leq (n+1)^{|\mathcal{X}|} \max_{P:H(P)>R_n} 2^{-nD(P||Q)} \\ &= (n+1)^{|\mathcal{X}|} 2^{-n[\min_{P:H(P)>R_n} D(P||Q)]} \end{aligned} \tag{356}$$

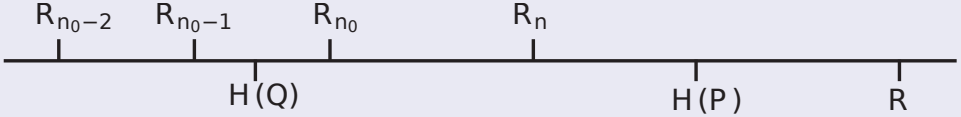
...

Teorema da Codificação de Shannon VII

Demonstração.

continuação...

- Temos que R_n forma uma sequência crescente com n, tal que $R_n < R$ para todo n.
- Por hipótese, $H(Q) < R$.
- Eventualmente, para algum n_0, teremos que $\forall n > n_0$, $R_n > H(Q)$.
- Na equação anterior, escolhemos $P : H(P) > R_n$.



- Teremos então: $H(P) > R_n > H(Q)$, o que implica em $P \neq Q$.
- Desta forma, teremos $D(P \,||\, Q) > 0$ para P escolhido.

...

Teorema da Codificação de Shannon VIII

Demonstração.

continuação...

- Teremos assim

$$P_e^{(n)} \leq \underbrace{(n+1)^{|\mathcal{X}|}}_{\text{polinomial em } n} \underbrace{2^{-n[\min_{P:H(P)>R_n} D(P||Q)]}}_{\text{exp. decrescente qnd } n\to\infty} \tag{357}$$

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

□

- Por outro lado, se $R < H(Q)$ teremos $P_e^{(n)} \to 1$.
- Entropia é o limite de compressão.

Processo Estocástico I

- Falamos de variáveis aleatórias i.i.d. $X_1, X_2, \ldots$. Neste contexto, cada uma delas possui a mesma entropia associada.
- O que ocorre quando as v.a.s. não são mais independentes? Como podemos lidar com a entropia do processo, neste caso?

Definição (Processo Estocástico Estacionário (sentido-estrito))

Uma sequência de v.a.s. $X_1, X_2, \ldots, X_n$ é governada por uma distribuição probabilística é dita estacionária em sentido estrito se

$$p(X_{1:n} = x_{1:n}) = p(X_{1+l:n+l} = x_{1:n}) \tag{358}$$

para todo l, todo n e todo $x_{1:n} \in \mathcal{X}^n$.

Processo de Markov I

Definição (Processo de Markov de primeira ordem)

Um processo estocástico é um processo de Markov de primeira ordem se

$$p(X_{n+1} = x_{n+1} \mid X_{1:n} = x_{1:n}) = p(X_{n+1} = x_{n+1} \mid X_n = x_n) \tag{359}$$

Neste caso, isto significa que $p(x_{1:n}) = p(x_1)p(x_2 \mid x_1) \dots p(x_n \mid x_{n-1})$.
Dado o presente, o futuro e o passado são independentes.

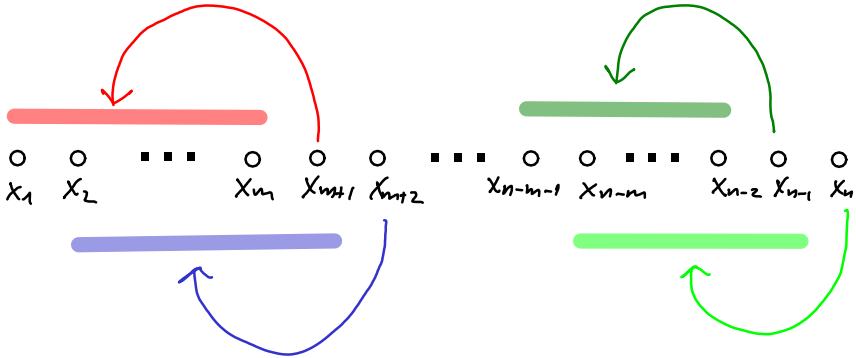
Definição (Processo de Markov de ordem m)

Um processo estocástico é um processo de Markov de ordem m se

$$p(X_{n+1} = x_{n+1} \mid X_{1:n} = x_{1:n}) = \\ p(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \dots, X_{n-m} = x_{n-m}) \tag{360}$$

Processo de Markov II

Neste caso, isto significa que $p(x_{1:n}) = p(x_{m+1} \mid x_m, x_{m-1}, \ldots, p(x_1)) \ldots p(x_{n-1} \mid x_{n-2}, x_{n-3}, \ldots, x_{n-m-1}) p(x_n \mid x_{n-1}, x_{n-2}, \ldots, x_{n-m})$.



Processo de Markov III

Definição (Homogêneo)

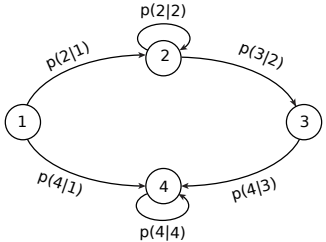
Uma cadeia de Markov é invariante no tempo (também chamada de homogênea) se $p(x_{n+1} \mid x_n)$ não depender do tempo, i.e., se

$$p(X_{n+1} = b \mid X_n = a) = p(X_2 = b \mid X_1 = a) \quad \forall a, b, n \tag{361}$$

Neste caso, a cadeia de Markov pode ser descrita por uma matriz de transição fixa $P = [p_{ij}]_{ij}$ em que $p_{ij} = p(X_{n+1} = j \mid X_n = i)$. Podemos representar esta cadeia de Markov como um grafo com setas entre estados cuja probabilidade de transição não é nula.

Processo de Markov IV

$$P = \begin{bmatrix} 0 & p(2 \mid 1) & 0 & p(4 \mid 1) \\ 0 & p(2 \mid 2) & p(3 \mid 2) & 0 \\ 0 & 0 & 0 & p(4 \mid 3) \\ 0 & 0 & 0 & p(4 \mid 4) \end{bmatrix} \quad (362)$$



- A probabilidade de um estado no instante $n+1$, dada em função dos possíveis estados no instante n e da probabilidade de transição:

$$p(x_{n+1}) = \sum_{x_n} p(x_n) p_{x_n x_{n+1}} \quad (363)$$

- uma cadeia de Markov de primeira ordem é estacionária se $p(x_{n+1}) = p(x_n)$.

Processo de Markov V

Definição (Irredutível)

Uma cadeia de Markov é irredutível se $p_{ij}(n) > 0$ para todo i, j e para algum n onde $p_{ij}(n) = p(X_{n+1} = j \mid X_n = i)$.

Ou seja, qualquer estado é acessível de qualquer outro estado (ao menos em algum instante n), com probabilidade não nula.

Processo de Markov VI

Definição (Período)

Uma cadeia de Markov é periódica se $d(i) > 1$ com

$$d(i) = \gcd\{n : p_{ii}(n) > 0\} \tag{364}$$

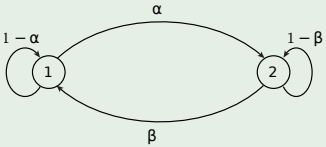
$d(i)$ é o período do i-ésimo estado.

Note que temos o máximo divisor comum do número de épocas para o qual um retorno ao mesmo estado é possível.
No caso de uma cadeia de Markov homogênea (invariante no tempo), se houver algum $p_{ii} > 0$, o período será 1 época.

Processo de Markov VII

Exemplo

$$P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \qquad (365)$$



- Se $\mu = [p_1 \; p_2]^T$ é uma distribuição estacionária, então devemos ter $\mu^T P = \mu^T$.
- Neste caso, teremos

$$\begin{aligned} \mu^T = \begin{pmatrix} p_1 & p_2 \end{pmatrix} &= \begin{pmatrix} p_1 & p_2 \end{pmatrix} \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \\ &= \begin{pmatrix} (1-\alpha)p_1 + \beta p_2 & \alpha p_1 + (1-\beta)p_2 \end{pmatrix} \end{aligned} \qquad (366)$$

...

Processo de Markov VIII

Exemplo

continuação...

Teremos então:

$$p_1 = (1-\alpha)p_1 + \beta p_2 \tag{367}$$

logo, $p_1 = \frac{\beta}{\alpha}p_2$ Sabemos também que devemos ter $p_1 + p_2 = 1$. Por conseguinte, teremos

$$\begin{aligned} p_1 + p_2 &= 1 \\ \frac{\beta}{\alpha}p_2 + p_2 &= 1 \\ p_2 &= \frac{\alpha}{\alpha+\beta} \end{aligned} \tag{368}$$

...

Processo de Markov IX

Exemplo

continuação...

Assim, teremos

$$\mu = \begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix} \tag{369}$$

Processo de Markov X

- Processo estocástico estacionário: a probabilidade dos estados não muda ao longo do tempo.
- Homogêneo: a matriz P de transições não muda ao longo do tempo.
- Processo de Markov: futuro e passado são independentes dado o presente (ou então: o passado imediado é suficiente, não sendo relevante o passado distante).
- Irredutível: todos estados são acessíveis, eventualmente.
- Periódico: máximo divisor comum entre os intervalos em que o retorno a um estado é possível.

Média Cesáro I

- considere a sequência $\{a_n, n \geq 1\}$
- construa a sequência $\{b_n, n \geq 1\}$, onde $b_n = \frac{1}{n} \sum_{i=1}^{n} a_i$
- b_n é a média Cesáro de $\{a_n\}$

Lema (Média Cesáro)

Sejam a_n números reais, se $a_n \to a$ quando $n \to \infty$ e $b_n = \frac{1}{n} \sum_{i=1}^{n} a_i$, então $b_n \to a$ quando $n \to \infty$.

Média Cesáro II

Demonstração.

- Como $a_n \xrightarrow[n\to\infty]{} a$, para todo $\epsilon > 0$, existe N_ϵ tal que $|a_n - a| < \epsilon$ para todo $n > N_\epsilon$.

...

Média Cesáro III

Demonstração.

continuação...

- Para $n > N_\epsilon$ teremos

$$\begin{aligned} |b_n - a| &= |\frac{1}{n}\sum_{i=1}^{n} a_i - a| = |\frac{1}{n}\sum_{i=1}^{n} a_i - \frac{1}{n}\sum_{i=1}^{n} a| \\ &= |\frac{1}{n}\sum_{i=1}^{n}(a_i - a)| \leq \frac{1}{n}\sum_{i=1}^{n}|a_i - a| \\ &\quad \text{onde utilizamos a desigualdade triangular} \\ &= \frac{1}{n}\left(\sum_{i=1}^{N_\epsilon}|a_i - a| + \sum_{i=N_\epsilon+1}^{n}|a_i - a|\right) \end{aligned} \tag{370}$$

...

Média Cesáro IV

Demonstração.

continuação...

$$
\begin{aligned}
|b_n - a| &\leq \frac{1}{n}\left(\sum_{i=1}^{N_\epsilon}|a_i - a| + \sum_{i=N_\epsilon+1}^{n}|a_i - a|\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{N_\epsilon}|a_i - a| + \frac{1}{n}\sum_{i=N_\epsilon+1}^{n}\epsilon \\
&= \frac{1}{n}\sum_{i=1}^{N_\epsilon}|a_i - a| + \underbrace{\frac{n - N_\epsilon}{n}}_{<1}\epsilon
\end{aligned} \tag{371}
$$

...

Média Cesáro V

Demonstração.

continuação...

$$
\begin{aligned}
|b_n - a| &\leq \frac{1}{n}\sum_{i=1}^{N_\epsilon}|a_i - a| + \underbrace{\frac{n - N_\epsilon}{n}}_{<1}\epsilon \\
&< \underbrace{\frac{1}{n}\sum_{i=1}^{N_\epsilon}|a_i - a|}_{<\epsilon} + \epsilon < 2\epsilon \qquad (372)
\end{aligned}
$$

onde utilizamos o fato de que podemos tomar n grande suficiente de forma que $\frac{1}{n}\sum_{i=1}^{N_\epsilon}|a_i - a| < \epsilon$, pois trata-se de uma soma finita.
Então $b_n \to a$ quando $n \to \infty$. □

Processo Estocástico I

- Processos Estocástico possuem taxas de entropia, que intuitivamente representam o quantidade de informação nova, na média, que é fornecida pelo processo estocástico a cada instante.

Definição (Taxa de Entropia de um processo estocástico)

A taxa de entropia de um processo estocástico $\{X_i\}_i$ é definida como

$$H(\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{373}$$

quando existir.

Note que, quando as v.a.s são i.i.d. teremos

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_{1:n}) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = H(X_1) \tag{374}$$

Processo Estocástico II

A taxa de entropia pode ser vista como a entropia por símbolo para um dado processo estocástico, quando n cresce indefinidamente.

Processo Estocástico III

Exemplo

- Se as v.a.s são independentes mas não são identicamente distribuídas, teremos

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) = ? \tag{375}$$

o que pode não existir.

Processo Estocástico IV

Definição (Taxa de Inovação da Informação (Definição Alternativa para Taxa de Entropia))

Vamos assumir um processo estocástico e definir a taxa da seguinte forma

$$H'(\mathcal{X}) \triangleq \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1) \tag{376}$$

se existir.

Veremos a seguir que $H'(\mathcal{X})$ existe para um processo estocástico estacionário.

Processo Estocástico V

Teorema

Para um processo estocástico estacionário, $H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$ é decrescente com n e possui como limite $H'(\mathcal{X})$.

Demonstração.

$$\begin{aligned} H(X_{n+1} \mid X_1, \ldots, X_n) &\leq H(X_{n+1} \mid X_2, \ldots, X_n) \\ &= H(X_n|X_1, \ldots, X_{n-1}) \end{aligned} \tag{377}$$

Onde utilizamos o fato de que condicionar não aumenta (decresce ou não altera) a entropia; e utilizamos o fato de que o processo estocástico é estacionário.
Temos então uma sequência decrescente com limite inferior 0, logo, esta sequência possui um limite: $H'(\mathcal{X})$. □

Processo Estocástico VI

Teorema

Para um processo estocástico estacionários temos

$$\lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1) \triangleq H'(\mathcal{X}) = H(\mathcal{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{378}$$

Demonstração.

$$b_n = \frac{H(X_1, X_2, \ldots, X_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} \underbrace{H(X_i \mid X_{i-1}, \ldots, X_1)}_{=a_i} \tag{379}$$

como $a_n \to H'(\mathcal{X})$, teremos $b_n \to H'(\mathcal{X})$, mas por definição $b_n \to H(\mathcal{X})$. □

Processo Estocástico VII

- Note que para qualquer processo estacionário ergódico, temos os seguinte:

$$-\frac{1}{n}\log p(x_1,\ldots,x_n) \to H(\mathcal{X}) \tag{380}$$

- Podemos mostrar algo como a Propriedade da Equipartição Assintótica para processos deste tipo (Capítulo 16.8).

Taxa de Entropia para Cadeia de Markov I

A taxa de entropia para um cadeia de Markov de primeira ordem estacionária será dada da seguinte forma

$$
\begin{aligned}
H(\mathcal{X}) &= H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \\
&= \lim_{n\to\infty} H(X_n \mid X_{n-1}) \\
&\quad \text{dado que é Markov de 1a ordem} \\
&= H(X_2 \mid X_1) \quad \text{(estacionário)} \\
&= -\sum_{x_2,x_1} p(x_2, x_1) \log p(x_2 \mid x_1) = \sum_i \mu_i \left[-\sum_j p_{ij} \log p_{ij} \right]
\end{aligned}
$$

onde μ é a distribuição estacionária e p_{ij} a probabilidade de transição de i para j.

Taxa de Entropia para Cadeia de Markov II

- Para o exemplo anterior, teremos

$$H(\mathcal{X}) = H(X_2 \mid X_1) = \frac{\beta}{\alpha+\beta}H(\alpha) + \frac{\alpha}{\alpha+\beta}H(\beta). \tag{381}$$

Caminhada do Bêbado I

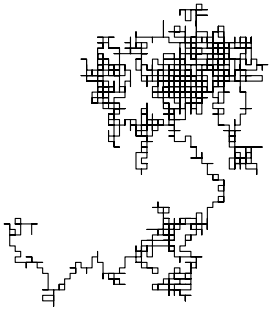


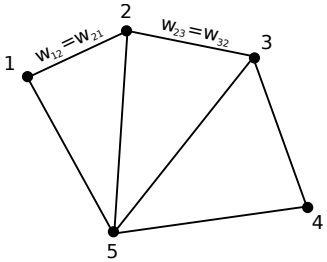
Figura 10: Passeio aleatório (Wikipedia).

Vamos considerar aqui o exemplo do passeio aleatório sobre um grafo com pesos.

- Assuma uma distribuição estacionária irredutível e aperiódica.

Caminhada do Bêbado II

- Considere o grafo $G = (V, E)$ com m nós rotulados $\{1, 2, \ldots, m\}$ e arestas entre os nós com pesos $w_{ij} \geq 0$ (aresta entre o nó i e o nó j). Teremos $w_{ij} = w_{ji}$ e $w_{ij} = 0$ se não existe aresta entre os nós i e j.



- O andar do bêbado (*random walk*) $\{X_n\}$, $X_n \in \{1, 2, \ldots, m\}$, é uma sequência de vértices de um grafo.

Caminhada do Bêbado III

- Dado $X_n = i$, o próximo vértice j será escolhido dentre aqueles nós conectados com i com probabilidade proporcional ao peso conectando os vértices i e j, i.e., p_{ij} será dado por

$$p_{ij} = \frac{w_{ij}}{\sum_j w_{ij}} = \frac{w_{ij}}{w_i} \tag{382}$$

  onde $w_i \triangleq \sum_j w_{ij}$ (peso total das arestas que saem do nó i).
- A soma de todos os pesos, de todas as arestas será dada por

$$w = \sum_{i,j:j>i} w_{ij} \tag{383}$$

- Note que $\sum_i w_i = \sum_{i,j} w_{ij} = 2w$.

Caminhada do Bêbado IV

- Vamos supor que a distribuição estacionária é dada por $\mu_i = \frac{w_i}{2w}$, o que poderemos checar verificando a equação $\mu^T = \mu^T P$:

$$\begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_m \end{pmatrix} = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_m \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1m} \\ p_{21} & p_{22} & \dots & p_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1} & p_{m2} & \dots & p_{mm} \end{pmatrix} \tag{384}$$

ou seja,

$$\forall j, \quad \sum_i \mu_i p_{ij} = \sum_i \frac{w_i}{2w} \frac{w_{ij}}{w_i} = \sum_i \frac{1}{2w} w_{ij} = \frac{w_j}{2w} = \mu_j \tag{385}$$

- $\mu_j = \frac{w_j}{2w}$: esta distribuição estacionária possui uma interessante propriedade de localização: ela depende apenas dos pesos locais (conectados ao nó em questão) e do total dos pesos; desta forma, ela não sofrerá alteração se os pesos de uma outra parte do grafo (com arestas não ligadas ao nó i) sofrerem alteração sem alterar a soma dos pesos.

Caminhada do Bêbado V

- Note que a cadeia é aperiódica, já que $w_{ii} = 0$, como evidenciado abaixo

$$\begin{aligned} 2w &= \sum_i w_i = \sum_{i,j} w_{ij} = \sum_{i,j:i=j} w_{i,j} + \sum_{i,j:i>j} w_{i,j} + \sum_{i,j:i<j} w_{i,j} \\ &= \sum_{i,j:i=j} w_{i,j} + w + w = \sum_{i,j:i=j} w_{i,j} + 2w \end{aligned} \tag{386}$$

logo, $\sum_{i,j:i=j} w_{i,j} = 0 \Rightarrow w_{ii} = 0$.

Caminhada do Bêbado VI

- A taxa de entropia para este passeio aleatório será dada por

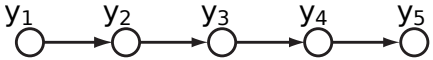
$$\begin{aligned} H(\mathcal{X}) &= H(X_2 \mid X_1) = -\sum_i \mu_i \sum_j p_{ij} \log p_{ij} \\ &= -\sum_i \frac{w_i}{2w} \sum_j \frac{w_{ij}}{w_i} \log \frac{w_{ij}}{w_i} = -\sum_{i,j} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{w_i} \\ &= -\sum_{i,j} \frac{w_{ij}}{2w} \log \left(\frac{w_{ij}}{w_i} \frac{2w}{2w} \right) \\ &= -\sum_{i,j} \frac{w_{ij}}{2w} \log \frac{w_{ij}}{2w} - \sum_{i,j} \frac{w_{ij}}{2w} \log \frac{2w}{w_i} \\ &= H\left(\ldots, \frac{w_{ij}}{2w}, \ldots \right) - H\left(\ldots, \frac{w_i}{2w}, \ldots \right) \end{aligned} \tag{387}$$

Caminhada do Bêbado VII

onde o primeiro termo representa a incerteza sob todas arestas e o segundo termo representa a incerteza sob todos os nós em condição estacionária ($w_i/2w = \mu_i$).
Se todas arestas possuírem o mesmo peso, sendo E_i o número de arestas emanando do nó i, e E o número total de arestas, teremos

$$H(\mathcal{X}) = \log(2E) - H\left(\frac{E_1}{2E}, \frac{E_2}{2E}, \ldots, \frac{E_m}{2E}\right). \tag{388}$$

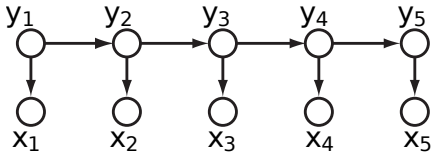
Modelo de Markov I



Para o modelo de Markov de primeira ordem, temos que dois estados são independentes quando um intermediário é dado, por exemplo:

$$Y_5 \perp\!\!\!\perp Y_3 \mid Y_4 \text{ ou então } Y_4 \perp\!\!\!\perp Y_1 \mid Y_{2:3} \text{ ou } Y_4 \perp\!\!\!\perp Y_1 \mid Y_2 \tag{389}$$

Cadeia Oculta Markov I



Na cadeia oculta de Markov, dizemos que um observação é gerada por um estado independente de outras observações ou estados passados. Por exemplo:

$$X_3 \perp\!\!\!\perp X_1 \mid Y_3 \tag{390}$$

Cadeia Oculta Markov II

- Seja $Y_1, Y_2, \ldots, Y_n$ uma cadeia de Markov estacionária.
- Seja $X_{1:n}$ uma função aleatória desta cadeia de Markov, i.e.,

$$X_i = \phi_N(Y_i) = \begin{cases} \phi_1(Y_i) & \text{com prob. } p_1(Y_i) \\ \phi_2(Y_i) & \text{com prob. } p_2(Y_i) \\ \vdots & \\ \phi_m(Y_i) & \text{com prob. } p_m(Y_i) \end{cases} \tag{391}$$

  onde $N \in \{1, 2, \ldots, m\}$ é uma variável aleatória.
- Note que o processo estocástico $X_1, X_2, \ldots$ não forma uma cadeia de Markov. Sequer a Markovidade de primeira ordem é satisfeita. Por exemplo: não podemos falar que $X_4 \perp\!\!\!\perp X_1 \mid X_{2:3}$, mesmo conhecendo $X_{2:3}$, X_1 ainda pode ser necessário para determinar X_4.
- Se $\{Y_i\}_i$ é estacionário, então $\{X_i\}_i$ é um processo estacionário.

Cadeia Oculta Markov III

- A taxa de entropia do processo $\{X_i\}_i$ pode ser calculada

$$H(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \tag{392}$$

mas iremos calcular os limites inferior e superior, o que é mais simples.

- Limite superior:

$$\begin{aligned} H(X_n \mid X_{n-1}, \ldots, X_1) &= H(X_{n+1} \mid X_n, \ldots, X_2) \\ &\geq H(X_{n+1} \mid X_n, \ldots, X_2, X_1) \\ &= H(X_{n+2} \mid X_{n+1}, \ldots, X_2) \\ &\geq H(X_{n+2} \mid X_{n+1}, \ldots, X_2, X_1) \\ &\geq \ldots \geq H(\mathcal{X}) \end{aligned} \tag{393}$$

Cadeia Oculta Markov IV

- Limite inferior:

$$
\begin{aligned}
H(X_n \mid X_{n-1}, \ldots, X_2, Y_1) &= H(X_n \mid X_{n-1}, \ldots, X_2, X_1, Y_1) \\
&\quad \text{já que } X_n \perp\!\!\!\perp X_1 \mid Y_1 \\
&= H(X_n \mid X_{n-1}, \ldots, X_2, X_1, Y_1, Y_0, Y_{-1}, \ldots, Y_{-k}) \qquad (394) \\
&\quad \text{já que } X_n \perp\!\!\!\perp Y_0 \mid Y_1 \text{ e } X_n \perp\!\!\!\perp Y_{-1} \mid Y_1 \text{ etc...} \\
&= H(X_n \mid X_{n-1}, \ldots, X_2, X_1, Y_1, Y_0, Y_{-1}, \ldots, Y_{-k}, X_0, \ldots, X_{-k}) \\
&\quad \text{pelo mesmo motivo} \\
&\leq H(X_n \mid X_{n-1}, \ldots, X_2, X_1, X_0, \ldots, X_{-k}) \\
&\quad \text{condicionar reduz a entropia} \\
&= H(X_{n+k+1} \mid X_{n+k}, \ldots, X_1) \qquad (395) \\
&\quad \text{estacionariedade}
\end{aligned}
$$

Cadeia Oculta Markov V

Desta forma temos o limite inferior:

$$H(X_n \mid X_{n-1}, \ldots, X_2, Y_1) \leq H(\mathcal{X}) \tag{396}$$

- Os limites para a taxa de informação em uma HMM são dados

$$H(X_n \mid X_{n-1}, \ldots, X_1, Y_1) \leq H(\mathcal{X}) \leq H(X_n \mid X_{n-1}, \ldots, X_1) \tag{397}$$

Cadeia Oculta Markov VI

Teorema (teorema do confronto (sanduíche))

$$H(X_n \mid X_{n-1}, \ldots, X_1) - H(X_n \mid X_{n-1}, \ldots, X_1, Y_1) \to 0 \tag{398}$$

Demonstração.

$$\begin{aligned} &H(X_n \mid X_{n-1}, \ldots, X_1) - H(X_n \mid X_{n-1}, \ldots, X_1, Y_1) \\ &\quad = I(X_n; Y_1 \mid X_{n-1}, \ldots, X_1) \leq H(Y_1) \end{aligned} \tag{399}$$

Temos também que $I(Y_1; X_1, \ldots, X_n) \leq H(Y_1)$, para todo n. ...

Cadeia Oculta Markov VII

Demonstração.

continuação...

Utilizando a regra da cadeira da informação mútua ($I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$), e tirando o limite, teremos

$$\lim_{n\to\infty} I(Y_1; X_1, \ldots, X_n) = \lim_{n\to\infty} \sum_{i=1}^{n} I(Y_1; X_i \mid X_{1:i-1})$$

$$= \sum_{i=1}^{\infty} I(Y_1; X_i \mid X_{1:i-1}) \leq H(Y_1) < \infty \tag{400}$$

Então o resultado da soma infinita é uma constante, isto significa que os termos $\to 0$ quando $n \to \infty$. Logo, cada um dos termos $I(Y_1; X_i \mid X_{1:i-1}) \to 0$ quando $n \to \infty$. □

Cadeia Oculta Markov VIII

Ao final, temos que

$$\lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1, Y_1) = H(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1) \tag{401}$$

Codificação I

- Codificação sem perdas de informação de fontes regidas por uma determinada distribuição.
- Teorema da Codificação de Shannon: é possível codificar utilizando $R > H(X)$ bits por símbolo da fonte, se utilizarmos bloco grande suficiente.
- Transmissão de blocos não é prática.
- Outras alternativas:
  1) Códigos de tamanho variável: os símbolos são codificados separadamente utilizando palavras de código com comprimento variável. Ex: Codificação Huffman.
  2) Codificação de Fluxo (*stream code*): codificação que opera no fluxo de dados, decidindo a palavra código dependendo do símbolo corrente e da história (símbolos passados). Ex: Codificação Aritmética, Codificação Lempel-Ziv (sendo este ultimo um codificação universal, pois não requer $p(x)$).
- Queremos uma codificação prática que alcance o limite da entropia.
- A codificação pode utilizar a distribuição $p(x)$ dada, ou estimá-la de alguma forma.
- Não iremos lidar aqui com o problema de estimar $p(x)$ (problema de estimação de densidade), vamos supor que $p(x)$ é conhecido ou sua aproximação $q(x)$ é dada.

Codificação II

- Vamos verificar o efeito de utilizar a aproximação $q(x)$ quando a distribuição subjacente é $p(x)$.

Codificação de Fonte I

Definição (Codificação de Fonte)

Um código C para a v.a. X é um mapeamento de $\mathcal{X}$ em $\mathcal{D}^*$, ou seja

$$C : \mathcal{X} \to \mathcal{D}^*, \tag{402}$$

onde $\mathcal{D}^*$ é o conjunto de sequências (*strings*) finitas em um alfabeto D-ário. $C(x)$ é a palavra código (*codeword*) correspondente ao símbolo x, e $l(x)$ é o comprimento desta palavra.

Exemplo

Seja $\mathcal{X} = \{\text{azul}, \text{vermelho}\}$. O código pode ser $C(\text{vermelho}) = 00$ e $C(\text{azul}) = 11$, o que seria um código binário para $\mathcal{D} = \{0, 1\}$. Teríamos então $\mathcal{D}^* = \{00, 11\}$. Uma sequência produzida pela fonte da forma $(\text{azul}, \text{azul}, \text{vermelho})$ seria codificada como $(11, 11, 00)$, ou melhor, 111100.

Codificação de Fonte II

Definição (Comprimento Esperado)

O comprimento esperado $L(C)$ de um código C para uma v.a. X com distribuição $p(x)$ é dado por

$$L(C) = \sum_x p(x)l(x) \tag{403}$$

Codificação de Fonte III

- De forma geral, $\mathcal{D} = \{0, 1, \ldots, D-1\}$, mas usualmente utilizamos $D = 2$.

Exemplo

Seja $\mathcal{X} = \{1, 2, 3, 4\}$ e $\mathcal{D} = \{0, 1\}$. Podemos definir o código através da tabela.

x	p(x)	c(x)	l(x)
1	$1/2$	0	1
2	$1/4$	10	2
3	$1/8$	110	3
4	$1/8$	111	3

...

Codificação de Fonte IV

Exemplo

continuação...

- Neste caso, teremos $H(X) = 1.75$ e $L(C) = El(X) = 1.75$, então este código é muito bom.
- A decodificação para este código é fácil.
- Qual é a sequência de símbolos que produziu a seguinte sequência 010110111111110100? R: 1,2,3,4,4,3,2,1.
- Com pontuação separando os símbolos: 0,10,110,111,111,110,10,0. Dizemos que o código possui pontuação automática.

Codificação de Fonte V

Exemplo

Vamos considerar $\mathcal{X} = \{1, 2, 3\}$ e $\mathcal{D} = \{0, 1\}$.

Código:

x	1	2	3
$p(x)$	$1/3$	$1/3$	$1/3$
$C(X)$	0	10	11

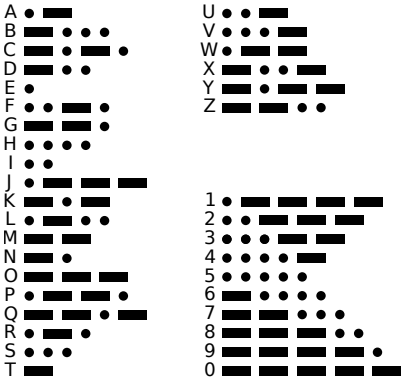
Teremos então $H = 1.58$bits e $El(X) = 1.66 > H$ bits.

Podemos facilmente decodificar? Por exemplo: $10110010 = 2, 3, 1, 1, 2$.

Codificação de Fonte VI

International Morse Code

1. The length of a dot is one unit.
2. A dash is three units.
3. The space between parts of the same letter is one unit.
4. The space between letters is three units.
5. The space between words is seven units.



Tipos de Código I

Definição (extensão)

Um extensão C^* de um código C é um mapeamento de uma sequência finita de símbolos em $\mathcal{X}$ em uma sequência finita de símbolos em $\mathcal{D}$ através da concatenação dos códigos:

$$C^*(x_1, x_2, \ldots, x_n) = C(x_1)C(x_2)\ldots C(x_n) \tag{404}$$

Exemplo

Se $C(x_1) = 0$ e $C(x_2) = 1$, então $C(x_1, x_2) = 01$.

Exemplo

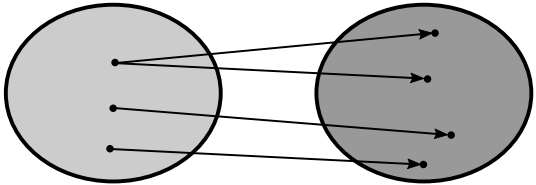
Se $C(x_1) = 000$ e $C(x_2) = 100$, então $C(x_1, x_2, x_2) = 000100100$.

Tipos de Código II

Definição (não-singular)

Um código é dito **não-singular** se todo elemento de $\mathcal{X}$ é mapeado em sequências (*strings*) diferentes em $\mathcal{D}^*$. I.e.,

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j) \tag{405}$$



Tipos de Código III

Definição (decodificação unívoca)

Um código C com extensão C^* é decodificável univocamente se a sua extensão C^* for não-singular.

- Objetivo: transmitir ou armazenar uma sequência de símbolos na forma de uma sequência de palavras codificadas.
- Um código não-singular pode se tornar único se inserirmos pontuação, mas para isto seria necessário incluir um novo símbolo no alfabeto $\mathcal{D}$, o que aumentaria a taxa R.
- É preferível um código com característica de auto-pontuação e que seja instantâneo.

Tipos de Código IV

Considere o código

x	1	2	3	4
$C(x)$	10	00	11	110

- Este código é unicamente decodificável.
- Ex.: 1100000000 = 3,2,2,2,2.
- Ex.: 11000000000 = 4,2,2,2,2.
- Note que só sabemos a identidade do primeiro símbolo após ler toda a sequência.

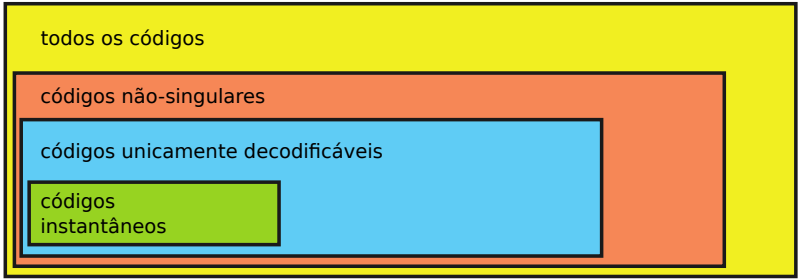
Tipos de Código V

Definição (código de prefixo)

Um código é chamado **código de prefixo** ou **código instantâneo** se nenhum palavra é prefixo de qualquer outra palavra.

- Quando temos um código de prefixo, sabemos onde está o fim de uma palavra, pois ela não é prefixo de nenhuma outra, ou seja, não existe nenhuma outra palavra que comece com a palavra encontrada.
- Um código de prefixo possui então a propriedade de auto-pontuação.
- Código de prefixo $\Rightarrow$ unicamente decodificável. Mas unicamente decodificável $\not\Rightarrow$ código de prefixo.

Tipos de Código VI



- Queremos um código com comprimento esperado mínimo possível.
- Analisando as classes de códigos, intuitivamente pensamos que é mais provável encontrarmos um código com comprimento esperado menor em uma classe maior (ou mais abrangente).
- Podemos alcançar um resultado melhor que a entropia se não utilizarmos um código não-singular, entretanto, queremos uma codificação sem perda.

Desigualdade de Kraft I

Teorema (Desigualdade de Kraft)

Para qualquer código instantâneo (código de prefixo) sobre um alfabeto de tamanho D, o comprimento das palavras $l_1, l_2, \ldots, l_m$ deve satisfazer

$$\sum_i D^{-l_i} \leq 1. \tag{406}$$

Por um outro lado, dado um conjunto de comprimentos de código satisfazendo a desigualdade acima, então existe um código de prefixo com estes comprimentos.

- Note que foi dito que existe um código de prefixo com aqueles comprimentos, não significa que todos códigos cujos comprimentos satisfazem a desigualdade são códigos de prefixo.

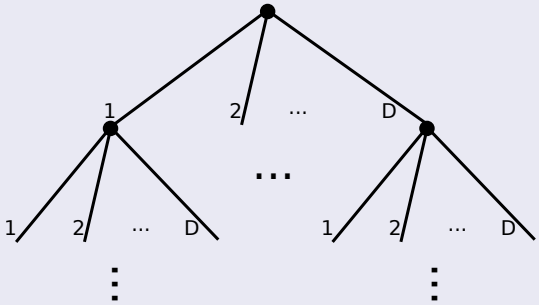
Desigualdade de Kraft II

- Se existe um código não-instantâneo com comprimentos l_i satisfazendo a desigualdade de Kraft, então podemos encontrar um outro código, que terá a propriedade de prefixo, e que possuirá os mesmos comprimentos l_i, consequentemente não alterando o comprimento esperado.
- Logo, será sempre melhor escolher um código de prefixo.

Desigualdade de Kraft III

Desigualdade de Kraft.

Vamos representar o conjunto de códigos em uma árvore D-ária (não necessariamente balanceada).



...

Desigualdade de Kraft IV

Desigualdade de Kraft.

continuação...

- As palavras correspondem à folhas na árvore.
- O caminho da raiz até a folha determina a palavra código.
- A condição de prefixo implica que não existe uma palavra a não ser nas folhas (nenhum descendente de uma palavra código será também uma palavra código).
- $l_{\max} = \max_i(l_i)$ é o comprimento da palavra mais longa.
- Podemos expandir toda a árvore até o comprimento $l_{\max}$.
- Os nós no nível de $l_{\max}$ são:
  1) palavras de código;
  2) descendentes de palavras de código; ou
  3) nem um, nem outro.

...

Desigualdade de Kraft V

Desigualdade de Kraft.

continuação...

- Considere uma palavras i no nível l_i da árvore (então esta palavra possui comprimento l_i).
- Existem então $D^{l_{\max}-l_i}$ descendentes de i, na árvore, no nível $l_{\max}$.
- Com a condição de prefixo, podemos afirmar que os descendentes de código i no nível l_i são disjuntos dos descendentes do código j no nível l_j, quando $i \neq j$ (i.e., o conjunto de descendentes para diferentes palavras é disjunto).
- O número total de nós em um conjunto de todos descendentes é $\leq D^{l_{\max}}$.

...

Desigualdade de Kraft VI

Desigualdade de Kraft.

continuação...

- Utilizando o que vimos acima, temos que a soma do número de descendentes de todos os códigos é menor ou igual ao número de folhas na árvore cheia no nível $l_{\max}$. Podemos então escrever

$$\sum_i D^{l_{\max}-l_i} \leq D^{l_{\max}} \quad \Rightarrow \quad \sum_i D^{-l_i} \leq 1 \tag{407}$$

- Por outro lado, dados os comprimentos $l_1, l_2, \ldots, l_m$, satisfazendo Kraft, iremos mostrar como construir um código de prefixo utilizando estes comprimentos.
- Considere uma árvore D-ária cheia de profundidade $l_{\max}$ e $D^{l_{\max}}$ nós terminais.

...

Desigualdade de Kraft VII

Desigualdade de Kraft.

continuação...

- Observe que
  1) No nível 0 existe uma fração 1 de descendentes de cada nó neste nível;
  2) No nível 1 existe uma fração $1/D$ de descendentes de cada nó neste nível;
  3) No nível 2 existe uma fração $1/D^2$ de descendentes de cada nó neste nível;
  4) ...
- De forma geral, em cada nível $i \in [0, l_{\max}]$ da árvore, existe uma fração de D^{-i} nós terminais que são descendentes de uma ramificação de cada um dos D^i nós no nível i.

...

Desigualdade de Kraft VIII

Desigualdade de Kraft.

continuação...

- Ordene então os comprimentos $(l_1, l_2, \ldots, l_m)$ de forma ascendente $(s_1, s_2, \ldots, s_m)$, sendo $s_1 \leq s_2 \leq \ldots \leq s_m$. Observe que existem tantos comprimentos quanto existem palavras.
- Para o comprimento s_1, escolha um nó no nível s_1 para indicar o código.
- Para garantir a condição de prefixo, o nó escolhido deve se tornar um nó terminal, eliminando assim uma fração D^{-s_1} de nós terminais no nível l_{max}.
- Em seguida, escolha um dos nós remanescentes no nível s_2 (neste momento, existem $(D^{s_1} - 1)D^{s_2 - s_1}$ possíveis escolhas), eliminando assim uma fração D^{-s_2} de nós terminais no nível l_{max}.

...

Desigualdade de Kraft IX

Desigualdade de Kraft.

continuação...

- A fração total de nós eliminados, até o momento, foi de $D^{-s_1} + D^{-s_2}$.
- Continuando este processo, iremos eliminar uma fração $\sum_{i=1}^{m} D^{-s_i}$ de nós, e neste processo estamos garantindo que estamos criando um código instantâneo (uma palavra de código não pode ser prefixo de outra).
- Como por suposição temos $\sum_{i=1}^{m} D^{-s_i} \leq 1$, nunca iremos eliminar mais do que todas as palavras de código, então este processo não irá exaurir as palavras de código.
- Criamos assim um código de prefixo com os comprimentos desejados, satisfazendo Kraft.

□

Kraft Infinito I

Teorema (Kraft infinito contável)

Para qualquer conjunto infinito contável de palavras de código que formam um conjunto com prefixo, este conjunto satisfaz a desigualdade de Kraft estendida, i.e.

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1 \tag{408}$$

Por outro lado, dados l_is satisfazendo a equação acima, existe um código de prefixo com estes comprimentos.

Kraft Infinito II

Kraft Infinito.

- Assuma que tenhamos um código de prefixo infinito contável e que o alfabeto é D-ário $\{0, 1, \ldots, D-1\}$.
- Considere a i-ésima palavra $y_1, y_2, \ldots, y_{l_i}$.
- A expansão da i-ésima palavra, de comprimento l_i, utilizando os dígitos fracionários é da forma

$$0.y_1y_2y_3\ldots y_{l_i} = \sum_{j=1}^{l_i} y_j D^{-j} \tag{409}$$

...

Kraft Infinito III

Kraft Infinito.

continuação...

- Considere os exemplos com $D = 2$ e $\mathcal{D} = \{0, 1\}$:
  - $0.1 = 1 \times 2^{-1} = \frac{1}{2}$
  - $0.01 = 0 \times 2^{-1} + 1 \times 2^{-2} = \frac{1}{4}$
  - $0.11 = 1 \times 2^{-1} + 1 \times 2^{-2} = \frac{3}{4}$
  - $0.001 = 1 \times 2^{-1} + 1 \times 2^{-2} + 1 \times 2^{-3} = \frac{1}{8}$
- Vamos associar cada palavra $y_{1:l_i}$ com o intervalo semiaberto na reta real $[0.y_1y_2 \dots y_{l_i}, 0.y_1y_2 \dots y_{l_i} + 1/D^{l_i})$.

...

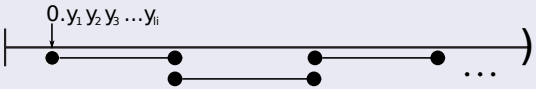
Kraft Infinito IV

Kraft Infinito.

continuação...

- Considere os exemplos com $D = 2$ e $\mathcal{D} = \{0, 1\}$:
  - A palavra $0.y_1 = 0.1 = \frac{1}{2}$ está associada ao intervalo $[\frac{1}{2}, 1)$.
  - A palavra $0.y_1y_2 = 0.01 = \frac{1}{4}$ está associada ao intervalo $[\frac{1}{4}, \frac{1}{2})$.
  - A palavra $0.y_1y_2y_3 = 0.001 = \frac{1}{8}$ está associada ao intervalo $[\frac{1}{8}, \frac{1}{4})$.

  Considere agora os exemplos seguintes com $D = 10$.
  - $0.y_1y_2y_3 = 0.157$ está associado ao intervalo $[0.157, 0.158)$.
  - $0.y_1y_2y_3 = 0.159$ está associado ao intervalo $[0.159, 0.160)$.



...

Kraft Infinito V

Kraft Infinito.

continuação...

- O intervalo da palavra $y_1 y_2 \dots y_{l_i}$ corresponde ao conjunto de todos números que iniciam-se com $y_1 y_2 \dots y_{l_i}$, e desta forma é um subintervalo do intervalo unitário.
- Além disso, $y_1 y_2 \dots y_{l_i}$ não é prefixo para nenhuma outra palavra, desta forma os intervalos serão disjuntos.
- O tamanho do intervalo associado à palavra $y_1 y_2 \dots y_{l_i}$ é D^{-l_i}.
- Como todos intervalos estão dentro do intervalo $[0, 1)$ teremos que

$$\sum_i D^{-l_i} \leq 1 \tag{410}$$

- A prova da proposição inversa é similar ao caso finito.

□

Em busca de um código ótimo I

- Código de prefixo $\Leftrightarrow$ desigualdade de Kraft.
- Precisamos encontrar os comprimentos l_i que satisfazem Kraft para obter um código de prefixo.
- Objetivo: encontrar um código de prefixo com o comprimento esperado mínimo.

$$L(C) = \sum_i p_i l_i \tag{411}$$

- Temos então um problema de otimização com restrição: encontrar

$$\min_{\{l_{1:m}\} \in \mathbb{Z}_+^m} \sum_i p_i l_i \tag{412}$$

sujeito a

$$\sum_i D^{-l_i} \leq 1 \tag{413}$$

Em busca de um código ótimo II

- Temos um problema de programação em inteiros que é um problema NP-difícil. Este problema não provável de ser solucionado eficientemente (a menos que $P = NP$).
- Vamos relaxar a condição de que l_i precisa ser inteiro e considerar o Lagrangiano

$$J = \sum_i p_i l_i + \lambda \left(\sum_i D^{-l_i} - 1 \right) \tag{414}$$

- Tomando as derivadas e igualando a zero, teremos:

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \ln D = 0$$

$$\Rightarrow D^{-l_i} = \frac{p_i}{\lambda \ln D} \tag{415}$$

$$\frac{\partial J}{\partial \lambda} = \sum_i D^{-l_i} - 1 = 0 \Rightarrow \lambda = 1/\ln D \tag{416}$$

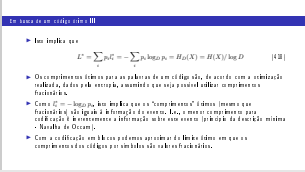
$$\Rightarrow D^{-l_i} = p_i \quad \text{implicando em} \quad l_i^* = -\log_D p_i \tag{417}$$

Em busca de um código ótimo III

- Isto implica que

$$L^* = \sum_i p_i l_i^* = -\sum_i p_i \log_D p_i = H_D(X) = H(X)/\log D \tag{418}$$

- Os comprimentos ótimos para as palavras de um código são, de acordo com a otimização realizada, dados pela entropia, assumindo que seja possível utilizar comprimentos fracionários.
- Como $l_i^* = -\log_D p_i$, isto implica que os "comprimentos" ótimos (mesmo que fracionários) são iguais à informação do evento. I.e., o menor comprimento para codificação é inerentemente a informação sobre este evento (princípio da descrição mínima - Navalha de Occam).
- Com a codificação em blocos podemos aproximar do limite ótimo em que os comprimentos dos códigos por símbolos são valores fracionários.



NP-Difícil

"Um problema H é NP-difícil se e somente se (sse) existe um problema NP-completo L que é Turing-redutível em tempo polinomial para H (i.e., L?=?TH). Em outras palavras, L pode ser resolvido em tempo polinomial por uma Máquina de Turing não determinística com um oráculo para H. Informalmente, podemos pensar em um algoritmo que pode chamar tal Máquina de Turing Não-Determinística como uma sub-rotina para resolver H, e resolver L em tempo polinomial, se a chamada da sub-rotina leva apenas um passo para computar. Problemas NP-difíceis podem ser de qualquer tipo: problemas de decisão, problemas de pesquisa ou problemas de otimização." (Wikipedia)

Em busca de um código ótimo III

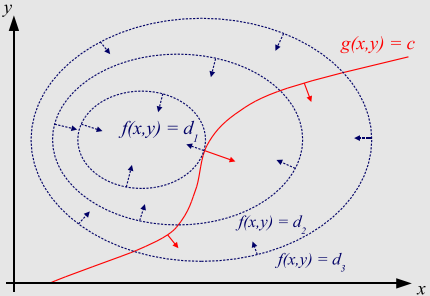
- Isso implica que

$$L^* = \sum_x p_x l_x^* = -\sum_x p_x \log_D p_x = H_D(X) = H(X)/\log D \qquad (4.11)$$

- Os comprimentos ótimos para as palavras de um código são, de acordo com a otimização realizada, dados pela entropia, a quantidade que seja possível utilizar os comprimentos fracionários.
- Como $l_x^* = -\log_D p_x$, isso implica que os "comprimentos" ótimos (mesmo que fracionários) são ligados à informação de eventos. I.e., o mesmo comprimento para codificação de eventos com a informação subjetiva com o princípio da Navalha de Occam).
- Com a codificação em blocos podemos aproximar os limites ótimos em que os comprimentos dos códigos por símbolos são valores factíveis.

"Em problemas de otimização, o método dos multiplicadores de Lagrange permite encontrar extremos (máximos e mínimos) de uma função de uma ou mais variáveis suscetíveis a uma ou mais restrições. Por exemplo, considere o problema de otimização:

- maximize $f(x, y)$ ou seja, deseja-se encontrar o ponto máximo desta função
- sujeito a $g(x, y) = c$.



Em busca de um código ótimo III

- Isto implica que

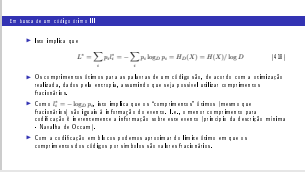
$$L^* = \sum_i p_i l_i^* = -\sum_i p_i \log_D p_i = H_D(X) = H(X)/\log D \tag{418}$$

- Os comprimentos ótimos para as palavras de um código são, de acordo com a otimização realizada, dados pela entropia, assumindo que seja possível utilizar comprimentos fracionários.
- Como $l_i^* = -\log_D p_i$, isto implica que os "comprimentos" ótimos (mesmo que fracionários) são iguais à informação do evento. I.e., o menor comprimento para codificação é inerentemente a informação sobre este evento (princípio da descrição mínima – Navalha de Occam).
- Com a codificação em blocos podemos aproximar do limite ótimo em que os comprimentos dos códigos por símbolos são valores fracionários.

O método consiste em introduzir uma variável nova (λ normalmente), chamada de multiplicador de Lagrange. A partir disso, estuda-se a função de Lagrange, assim definida:

$$\Lambda(x, y, \lambda) = f(x, y) + \lambda \cdot \Big(g(x, y) - c\Big), \tag{419}$$

Nesta função, o termo λ pode ser adicionado ou subtraído. Se $f(x, y)$ é um ponto de máximo para o problema original, então existe um λ tal que (x, y, λ) é um ponto estacionário para a função lagrangiana, ou seja, existe um ponto para o qual as derivadas parciais de Λ são iguais a zero.
No entanto, nem todos os pontos estacionários permitem uma solução para o problema original. Portanto, o método dos multiplicadores de Lagrange garante uma condição necessária para a otimização em problemas de otimização com restrição." (Wikipedia)



Navalha de Occam

"A Navalha de Occam ou Navalha de Ockham é um princípio lógico atribuído ao lógico e frade franciscano inglês Guilherme de Ockham (século XIV).

O princípio afirma que a explicação para qualquer fenómeno deve supor apenas as premissas estritamente necessárias à sua explicação e eliminar todas as que não causariam qualquer diferença aparente nas predições da hipótese ou teoria. O princípio é frequentemente designado pela expressão latina *Lex Parsimoniae* (Lei da Parcimónia) enunciada como: *"entia non sunt multiplicanda praeter necessitatem"* (as entidades não devem ser multiplicadas além da necessidade).

O princípio recomenda assim que se escolha a teoria explicativa que implique o menor número de premissas assumidas e o menor número de entidades.

Código Ótimo I

Teorema

Entropia é o comprimento esperado mínimo. O comprimento esperado L de qualquer código D-ário instantâneo (satisfaz Kraft por conseguinte) para uma v.a. X é tal que

$$L \geq H_D(X) \tag{420}$$

com igualdade sse $D^{-l_i} = p_i$.

Código Ótimo II

Demonstração.

$$\begin{aligned} L - H_D(X) &= \sum_i p_i l_i - \sum_i p_i \log_D 1/p_i \\ &= -\sum_i p_i \log_D D^{-l_i} + \sum_i p_i \log_D p_i \end{aligned} \tag{421}$$

...

Código Ótimo III

Demonstração.

continuação...

$$
\begin{aligned}
L - H_D(X) &= \dots \\
&= -\sum_i p_i \log_D D^{-l_i} + \\
&\quad \underbrace{\log_D \left(\sum_i D^{-l_i} \right) - \log_D \left(\sum_i D^{-l_i} \right)}_{=0} + \\
&\quad \sum_i p_i \log_D p_i \qquad (422)
\end{aligned}
$$

...

Código Ótimo IV

Demonstração.

continuação...

(interlúdio)

$$-\sum_i p_i \log_D D^{-l_i} + \log_D \left(\sum_j D^{-l_j} \right) + \sum_i p_i \log_D p_i =$$

$$-\sum_i p_i \log_D D^{-l_i} + \left(\underbrace{\sum_i p_i}_{=1} \right) \log_D \left(\sum_i D^{-l_i} \right) + \sum_i p_i \log_D p_i =$$

$$-\sum_i p_i \log_D D^{-l_i} + \sum_i p_i \log_D \left(\sum_j D^{-l_j} \right) + \sum_i p_i \log_D p_i = \ldots \quad (423)$$

...

Código Ótimo V

Demonstração.

continuação...

$$\begin{aligned} \ldots &= \sum_i p_i \log_D \frac{p_i \left(\sum_j D^{-l_j}\right)}{D^{-l_i}} \\ &= \sum_i p_i \log_D \frac{p_i}{\frac{D^{-l_i}}{\left(\sum_j D^{-l_j}\right)}} = \sum_i p_i \log_D \frac{p_i}{r_i} \end{aligned} \tag{424}$$

onde definimos

$$r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}. \tag{425}$$

...

Código Ótimo VI

Demonstração.

Código Ótimo VII

continuação...

Teremos assim

$$\begin{aligned} L - H_D(X) &= \dots \\ &= \sum_i p_i \log_D \frac{p_i}{r_i} - \log_D \left(\sum_i D^{-l_i} \right) \\ &= \underbrace{D(p \,||\, r)}_{\geq 0} + \underbrace{\log_D (1/c)}_{\geq 0} \\ &\geq 0 \end{aligned} \tag{426}$$

já que $c \leq 1$ pois satisfaz Kraft,

onde $c = \sum_i D^{-l_i}$.

□

Código Ótimo VIII

- Mostramos que $L \geq H_D(X)$.
- A igualdade $L = H$ é alcançada sse $p_i = D^{-l_i}$ para todo i sse $-\log_D p_i$ for um inteiro. Neste caso teremos $c = \sum_i D^{-l_i} = 1$.

Definição (D-ádico)

Uma distribuição probabilística é D-ádica com relação a D se cada uma das probabilidades é da forma D^{-n} para algum n.

- Exemplo: Quando $D = 2$, a distribuição $\left[\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right] = [2^{-1}, 2^{-2}, 2^{-3}, 2^{-3}]$ é 2-ádica.
- Teremos $L = H$ quando a distribuição for D-ádica.

Sumário I

- No problema de otimização relaxado, retiramos a restrição de que os comprimentos dos códigos precisam ser inteiros, desta forma obtivemos $El = H$.
- Assumindo que Kraft seja satisfeito (existe um código de prefixo com tais comprimentos) teremos comprimentos inteiros de forma que $El \geq H$.
  - Se ainda $l_i = -\log_D p_i$ é inteiro, para todo i, então teremos a igualdade $El = H$.
  - Se $l_i \neq -\log_D p_i$, mas l_i é inteiro, teremos estritamente $El > H$.

Como construir um código buscando o ótimo I

- Vimos anteriormente que o limite ótimo para o comprimento esperado do código é a entropia. Desta forma $L - H$ é uma medida do quão distante estamos deste ponto ótimo.
- $L - H = D(p \,||\, r) + \log_D 1/c$, onde $c = \sum_i D^{-l_i}$ e $r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}$.
- Para criar o código devemos buscar a distribuição D-ádica mais próxima (em termos de divergência de KL) da distribuição p e então construir um código seguindo a proposição inversa de Kraft.
- De forma geral, a menos que $P = NP$, é difícil encontrar a distribuição D-ádica mais próxima (no sentido de KL) da distribuição p (problema de programação linear).

Códigos de Shannon I

- Considerar o comprimento como inteiro mais próximo do valor ótimo: $l_i = \lceil \log_D 1/p_i \rceil$.
- Devemos verificar se esta escolha satisfaz Kraft:

$$\sum_i D^{-l_i} = \sum_i D^{-\lceil \log_D 1/p_i \rceil} \leq \sum_i D^{-\log_D 1/p_i} = \sum_i p_i = 1 \tag{427}$$

  Kraft é satisfeito, então existe um código de prefixo com estes comprimentos.
- Temos então os seguintes limites para os comprimentos:

$$\log_D \frac{1}{p_i} \leq l_i \leq \log_D \frac{1}{p_i} + 1 \tag{428}$$

Códigos de Shannon II

- Tomando o valor esperando teremos:

$$\begin{aligned} E_p\left[\log_D \frac{1}{p_i}\right] &\leq E_p\left[l_i\right] \leq E_p\left[\log_D \frac{1}{p_i} + 1\right] \\ \sum_i p_i \log_D \frac{1}{p_i} &\leq L \leq \sum_i p_i \left(\log_D \frac{1}{p_i} + 1\right) \\ H_D(X) &\leq L \leq H_D(X) + 1 \end{aligned} \tag{429}$$

  Ou seja, no máximo a 1 bit (ou D-it, bit se $D = 2$) de distância da entropia.

- Além disso temos que $H \leq L^* \leq L$, onde L^* é o comprimento ótimo para códigos de comprimento inteiro (obs.: é possível que os comprimentos ótimos inteiros não satisfaçam Kraft).

Códigos de Shannon III

Teorema

Sejam $l_1^*, l_2^*, \ldots, l_m^*$ comprimentos inteiros ótimos de códigos para uma fonte p e um alfabeto D-ário. L^* é o comprimento esperado. Então

$$H_D(X) \leq L^* \leq H_D(X) + 1 \tag{430}$$

- O custo adicional de utilizar inteiros (ao invés de valores fracionários) para os comprimentos das palavras não é maior do que um bit por símbolo.
- Este custo adicional é significativo? (muito ruim?) Depende de H.
- Definimos então a eficiência de um código:

Códigos de Shannon IV

Definição (Eficiência de um código)

A eficiência de um código é definida da seguinte forma

$$0 \leq \text{eficiência} \triangleq \frac{H_D(X)}{El(X)} \leq 1 \tag{431}$$

- Se $El(X) = H_D(X) + 1$, então a eficiência $\to 1$ quando $H(X) \to \infty$.
- Eficiência $\to 0$ quando $H(X) \to 0$.
- A entropia precisa ser muito grande para que a tenhamos uma boa eficiência.
- Para alfabetos pequenos é impossível ter boa eficiência. Por exemplo: $\mathcal{D} = \{0, 1\}$, então $\max H(X) = 1$, desta forma a melhor eficiência possível é 50%.

Melhorando a eficiência I

- O código visto anteriormente é inerentemente desvantajoso, a menos que a distribuição seja D-ádica.
- Podemos melhorar a eficiência codificando mais de um símbolo por vez (codificação em bloco). Isto equivale a aumentar o alfabeto e aumentar a entropia.
- Vamos chamar de L_n o comprimento esperado por símbolo de uma sequência de n símbolos $x_{1:n}$.

$$L_n = \frac{1}{n} \sum_{x_{1:n} \in \mathcal{X}^n} p(x_{1:n}) l(x_{1:n}) = \frac{1}{n} E l(x_{1:n}) \tag{432}$$

- Utilizando os comprimentos do código de Shannon, teremos

$$\begin{array}{rcl} \log 1/p_i & \leq l_i \leq & \log 1/p_i + 1 \\ \sum_i p_i \log 1/p_i & \leq \sum_i p_i l_i \leq & \sum_i p_i \left(\log 1/p_i + 1\right) \\ H(X_1, \ldots, X_n) & \leq E l(X_{1:n}) \leq & H(X_1, \ldots, X_n) + 1 \end{array} \tag{433}$$

Melhorando a eficiência II

- Se X_i são i.i.d., então $H(X_1, \ldots, X_n) = nH(X_i)$.
- Teremos assim

$$H(X) \leq L_n \leq H(X) + \frac{1}{n} \tag{434}$$

- Quando n cresce, a penalidade por símbolo imposta ao código de Shannon diminui e conseguiremos assim aproximar-nos do limite da Entropia (por símbolo), embora teremos que mais uma vez adotar a estratégia de codificação em blocos.

Processos Estocásticos I

- Considere um processo estocástico estacionário (ergódico), então

$$\begin{array}{lll} H(X_1,\ldots,X_n) & \leq El(X_{1:n}) \leq & H(X_1,\ldots,X_n)+1 \\ \dfrac{H(X_1,\ldots,X_n)}{n} & \leq L_n \leq & \dfrac{H(X_1,\ldots,X_n)}{n}+\dfrac{1}{n} \end{array} \tag{435}$$

- Se o processo é estacionário, então $L_n \to H(\mathcal{X})$ (taxa de entropia) quando $n \to \infty$.

Teorema

O comprimento mínimo esperado de código por símbolo satisfaz

$$\frac{H(X_1,\ldots,X_n)}{n} \leq L_n^* \leq \frac{H(X_1,\ldots,X_n)}{n}+\frac{1}{n} \tag{436}$$

se X_i é estacionário. Então $L_n^* \to H(\mathcal{X})$.

Codificando para a distribuição errada. I

- Em geral a distribuição subjacente não é conhecida. Isto implica que existem erros nos cálculos.
- O código de Shannon utiliza $l(x) = \lceil \log 1/q(x) \rceil$, mas a real distribuição é $p(x) \neq q(x)$. O que implica essa diferença?
- Vamos recalcular o valor esperado do comprimento do código, agora utilizando esta informação.

$$\begin{aligned} El(X) &= \sum_x p(x) \lceil \log 1/q(x) \rceil \leq \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right) \\ &= \sum_x p(x) \left(\log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1 \right) \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\ &= D(p \,||\, q) + H(p) + 1 \end{aligned} \tag{437}$$

Codificando para a distribuição errada. II

- $D(p \,||\, q)$ é o custo adicional por símbolo causado por utilizar a distribuição errada.

Teorema

O comprimento esperado de um código sob a distribuição $p(x)$ com $l(x) = \lceil \log 1/q(x) \rceil$ satisfaz

$$H(p) + D(p \,||\, q) \leq E_p l(X) \leq H(p) + D(p \,||\, q) + 1 \tag{438}$$

- No melhor caso estaremos sofrendo uma penalidade de $D(p \,||\, q)$.

Kraft revisitado I

- O objetivo é encontrar um código unívoco (unicamente decodificável) com comprimento esperado de palavra mínimo.
- Observando as classes de códigos, podemos imaginar que a classe mais ampla é a mais provável de se encontrar tal código desejado.

- Kraft é verdadeiro para códigos instantâneos (e vice-versa).

Kraft revisitado II

- Dentre os códigos unicamente decodificáveis, podemos encontrar um código melhor (menor comprimento esperado) do que um código de prefixo? (Já que estamos analisando uma classe maior de código).

Teorema

O comprimento de palavras de qualquer código unicamente decodificável (não necessariamente instantâneo) deve satisfazer a desigualdade de Kraft $\sum_i D^{-l_i} \leq 1$. *Por outro lado, dado um conjunto de comprimentos que satisfazem Kraft, é possível construir um código unicamente decodificável.*

Demonstração.

A proposição inversa já foi previamente mostrada, uma vez que mostramos como construir um código instantâneo utilizando um conjunto de comprimentos satisfazendo Kraft. □

Kraft revisitado III

Antes de provar o teorema vamos mostrar o seguinte Lemma.

Lema

Seja $\mathcal{X} = \{x_1, x_2, \ldots, x_m\}$, $|\mathcal{X}| = m$, temos o seguinte

$$\begin{aligned}
\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k &= \left(\sum_{i=1}^{m} D^{-l(x_i)}\right)^k \\
&= \left(D^{-l(x_1)} + D^{-l(x_2)} + \ldots + D^{-l(x_m)}\right)^k \\
&= \sum_{\substack{n_1,n_2,\ldots,n_m\geq 0 \\ n_1+n_2+\ldots+n_m=k}} \frac{k!}{n_1!n_2!\ldots n_m!} D^{-l(x_1)n_1}\ldots D^{-l(x_m)n_m}
\end{aligned}$$

...

Kraft revisitado IV

Lema

continuação...

Cada termo no somatório acima é devido a uma sequência de comprimento k formada pela combinação dos símbolos em $\mathcal{X}$.
Podemos interpretar a série como

$$\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k = \sum_{x_{1:k}\in\mathcal{X}^k} D^{-l(x_1)}D^{-l(x_2)}\dots D^{-l(x_k)} \tag{439}$$

Kraft revisitado V

Demonstração.

- Dado um código unicamente decodificável (não necessariamente instantâneo) com comprimentos $l(x)$ e com extensão com comprimentos dados por $l(x_1, \ldots, x_k) = \sum_{i=1}^{k} l(x_i)$, queremos mostrar que $\sum_x D^{-l(x)} \leq 1$.

...

Kraft revisitado VI

Demonstração.

continuação...

- Vamos definir $S = \sum_{x \in \mathcal{X}} D^{-l(x)}$ e então

$$\begin{aligned} S^k &= \left(\sum_{x \in \mathcal{X}} D^{-l(x)} \right)^k \\ &= \sum_{x_{1:k} \in \mathcal{X}^k} D^{-l(x_1)} D^{-l(x_2)} \dots D^{-l(x_k)} \\ &\quad \text{(veja lemma)} \\ &= \sum_{x_{1:k} \in \mathcal{X}^k} D^{-\left(\sum_{i=1}^k l(x_i)\right)} \end{aligned} \tag{440}$$

...

Kraft revisitado VII

Demonstração.

continuação...

$$S^k = \dots \tag{441}$$

$$= \sum_{x_{1:k} \in \mathcal{X}^k} D^{-l(x_{1:k})}$$

$$= \sum_{m=1}^{k l_{\max}} a(m) D^{-m} \tag{442}$$

onde $l_{\max} = \max_x l(x)$, $a(m)$ é o número de sequências $x_{1:k}$ mapeadas em palavras de comprimento m, i.e.,

$$a(m) = \left| \{ x_{1:k} \in \mathcal{X}^k : l(x_{1:k}) = m \} \right| \tag{443}$$

...

Kraft revisitado VIII

Demonstração.

continuação...

Existem D^m palavras de comprimento m, e cada uma delas pode ter no máximo uma sequência da fonte associada, já que o código é unicamente decodificável. Então $a(m) \leq D^m$, e assim

$$\begin{aligned} \underbrace{S^k}_{\text{exponencial em } k} &= \sum_{m=1}^{kl_{\max}} a(m) D^{-m} \\ &\leq \sum_{m=1}^{kl_{\max}} D^m D^{-m} = \underbrace{kl_{\max}}_{\text{polinomial em } k} \quad \forall k \end{aligned} \tag{444}$$

Isto só pode ser verdade para todo k se $S \leq 1$. ...

Kraft revisitado IX

Demonstração.

continuação...

Teremos então

$$S = \sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1. \tag{445}$$

□

Código unicamente decodificável vs Código Instantâneo

- Todo código instantâneo é unicamente decodificável.
- Todos códigos unicamente decodificável devem satisfazer Kraft.
- Então podemos utilizar os mesmos comprimentos de palavras e construir um código de prefixo.
- Conclusão: não faz sentido utilizar a classe mais ampla de código unicamente decodificáveis, é melhor utilizar um código de prefixo, que terá o mesmo comprimento esperado e será instantâneo.

Algoritmo de Sardinas-Patterson I

O algoritmo de Sardinas-Patterson fornece uma maneira de determinar, em tempo polinomial, se um determinado código é unicamente decodificável ou não (ver exercício 5.27 do livro).

Sardinas, August Albert; Patterson, George W. (1953), "A necessary and sufficient condition for the unique decomposition of coded messages", Convention Record of the I.R.E., 1953 National Convention, Part 8: Information Theory, pp. 104-108.

Em busca do código ótimo I

- Código de Prefixo $\leftrightarrow$ desigualdade de Kraft.
- Precisamos apenas encontrar os comprimentos l_i que satisfazem Kraft e então criar um código de prefixo com estes comprimentos.
- Objetivo: minimizar o comprimento esperado do código.

$$L(C) = \sum_i p_i l_i \tag{446}$$

- Problema de otimização com restrição:

$$\min_{\{l_{1:m}\} \in \mathbb{Z}_+^m} \sum_i p_i l_i \quad \text{sujeito a} \sum_i D^{-l_i} \leq 1 \tag{447}$$

- Problema de programação em inteiros que é um problema NP-difícil, provavelmente não será solucionado de forma eficiente (a menos que $P = NP$).

Em busca do código ótimo II

- Relaxar a condição sobre l_i ser inteiro. Considerar o Lagrangiano e encontrar os comprimentos ótimos.

$$l_i^* = -\log_D p_i \tag{448}$$

- Entropia é o comprimento esperado mínimo: $L \geq H_D(X)$.
- $L - H$ é uma medida do quão distante estamos do ótimo.

$$L - H = D(p \,||\, r) + \log_D 1/c, \tag{449}$$

onde $c = \sum_i D^{-l_i}$ e $r_i = \frac{D^{-l_i}}{\sum_j D^{-l_j}}$.

- Para construir um código, encontramos a distribuição D-ádica mais próxima (KL) de p e construímos um código seguindo a proposição inversa de Kraft.

Em busca do código ótimo III

- Código de Shannon: $l_i = \lceil \log_D 1/p_i \rceil$. Estes comprimentos satisfazem Kraft. Logo existe um código de prefixo com estes comprimentos. Teremos $H_D(X) \leq L \leq H_D(X) + 1$. Teremos que a eficiência $\to 1$ quando $H(X) \to \infty$, mas quando $H(X) \to 0$ teremos que a eficiência $\to 0$. Por exemplo, quando $D = 2$ teremos eficiência máxima de 50%.

$$0 \leq \text{eficiência} \triangleq \frac{H_D(X)}{El(X)} \leq 1 \tag{450}$$

- Podemos melhorar a eficiência codificado em blocos.

$$H(X_1, \ldots, X_n) \leq El(X_{1:n}) \leq H(X_1, \ldots, X_n) + 1 \tag{451}$$

- Codificar com a distribuição errada q implica em:

$$H(p) + D(p \,||\, q) \leq E_p l(X) \leq H(p) + D(p \,||\, q) + 1 \tag{452}$$

onde utilizamos $l(x) = \lceil \log 1/q(x) \rceil$ e a distribuição subjacente é p.

Código de Shannon é ótimo? I

Exemplo

- Suponha $\mathcal{X} = \{0, 1\}$ com $p(X = 0) = 10^{-1000} = 1 - p(X = 1)$.
- Comprimentos de Shannon serão:
  - $l(0) = \lceil \log_2 10^{-1000} \rceil = 3322$ bits.
  - $l(1) = \lceil \log_2 (1 - 10^{-1000}) \rceil = 1$ bit.

  Para o símbolo 0 estamos utilizando 3321 bits a mais do que o necessário.
- De forma geral, pode acontecer que $\lceil \log_D p_i \rceil$ é maior do que o necessário.
- O código de Shannon não é um código de prefixo com comprimentos inteiros ótimo.

Código de Huffman I

- **Procedimento** para encontrar o menor código de prefixo. (Por ser um procedimento é difícil analisar matematicamente).
- Objetivo: dado $p(x)$, queremos encontrar o menor código de prefixo.

Método Guloso I

- Abordagem gulosa: começar pelo topo e dividir as palavras potenciais em probabilidades iguais (i.e., realizar perguntas com entropia máxima).
  - Esta abordagem é similar ao jogo das 20 perguntas. Temos um conjunto de objetos $\mathbf{S} = \{1, 2, 3, \ldots, m\}$ que ocorre com frequência proporcional aos pesos não negativos $(w_1, w_2, \ldots, w_m)$.
  - Queremos determinar qual o objeto realizando o menor número de perguntas possível.
  - Cada pergunta será da forma '$X \in \mathbf{A}$?' onde $\mathbf{A} \subseteq \mathbf{S}$.
  - Supondo $\mathbf{S} = \{x_1, x_2, x_3, x_4, x_5\}$.

Método Guloso II

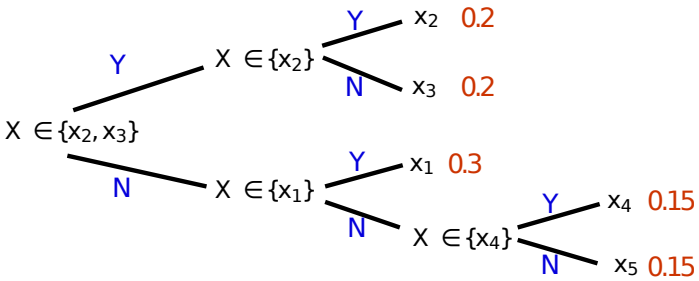


Figura 11: Exemplo (Bilmes, 2013).

- Charles Sanders Peirce, 1901

  Thus twenty skillful hypotheses will ascertain what two hundred thousand stupid ones might fail to do. The secret of the business lies in the caution which breaks a hypothesis up into its smallest logical components, and only risks one of them at a time.

- Método Guloso: na próxima etapa fazer aquilo que parece melhor naquele instante.

Método Guloso III

- Considere a distribuição

	a	b	c	d	e	f	g
p	0.01	0.24	0.05	0.20	0.47	0.01	0.02

- A pergunta que parece melhor é aquela que infere mais sobre a distribuição, reduz a entropia residual sobre X, será então a pergunta Y_1 com maior entropia. Teremos

$$H(X|Y_1) = H(X,Y_1) - H(Y_1) \quad (453)$$
$$= H(X) - H(Y_1) \quad (454)$$

onde utilizamos o fato de que Y_1 é uma função de X, e assim $H(X,Y_1) = H(X)$.

- Escolhemos a pergunta Y_1 com maior informação mútua com X.

$$I(Y_1;X) = H(X) - H(X|Y_1) = H(Y_1) \quad (455)$$

- As perguntas são da forma '$X \in \mathbf{A}$?' onde $\mathbf{A} \subseteq \mathbf{S}$. Desta forma, escolher uma pergunta do tipo sim-não é o mesmo que escolher o conjunto $\mathbf{A}$.

Método Guloso IV

- Considere a seguinte partição $\{a,b,c,d,e,f,g\} = \{a,b,c,d\} \cup \{e,f,g\}$. A pergunta '$X \in \{e,f,g\}$?' terá entropia máxima, pois $p(X \in \{a,b,c,d\}) = p(X \in \{e,f,g\}) = 0.5$.
- A pergunta corresponde à v.a. $Y_1 = \mathbf{1}_{\{X \in \{e,f,g\}\}}$, então $H(Y_1) = 1$, o que seria considerado uma boa pergunta.
- A próxima pergunta depende do resultado da pergunta anterior.
- Se $Y_1 = 0$ ($\equiv X \in \{a,b,c,d\}$) poderemos fazer a partição $\{a,b,c,d\} = \{a,b\} \cup \{c,d\}$, já que $p(\{a,b\}) = p(\{c,d\}) = 1/4$. Ou seja, $p(X \in \{a,b\} \mid X \in \{a,b,c,d\}) = 1/2$ e $p(X \in \{c,d\} \mid X \in \{a,b,c,d\}) = 1/2$.
- Esta pergunta corresponde à v.a. $Y_2 = \mathbf{1}_{\{X \in \{c,d\}\}}$. Teremos $H(Y_2 \mid Y_1 = 0) = 1$, e assim esta seria considerada uma boa pergunta.

Método Guloso V

- Para $Y_1 = 1$ precisaremos particionar $\{e, f, g\}$

	I	II	III
partição	$(\{e\}, \{f, g\})$	$(\{e, f\}, \{g\})$	$(\{e, g\}, \{f\})$
prob	$(0.47, 0.03)$	$(0.48, 0.02)$	$(0.49, 0.01)$
$H(Y_2 \mid Y_1 = 1)$	0.3274	0.2423	0.1414

A partição I é aquela que fornece a pergunta com maior entropia.

- Temos

$$\begin{aligned} H(X \mid Y_2, Y_1) &= H(X, Y_2 \mid Y_1) - H(Y_2 \mid Y_1) \\ &= H(X \mid Y_1) - H(Y_2 \mid Y_1) \\ &= H(X) - H(Y_2 \mid Y_1) - H(Y_1) \end{aligned} \tag{456}$$

- Abordagem gulosa: a cada passo escolhemos o que nos parece melhor, as escolhas futuras terão que lidar com as possibilidades que restaram devido às escolhas passadas.

Método Guloso VI

conjunto	partição	probabilidades	entropia condicional
$\{a,b,c,d,e,f,g\}$	$\{a,b,c,d\}$, $\{e,f,g\}$	$(0.5, 0.5)$	$H(Y_1) = 1$
$\{a,b,c,d\}$	$\{a,b\}$, $\{c,d\}$	$(0.25, 0.25)$	$H(Y_2 \mid Y_1 = 0) = 1$
$\{e,f,g\}$	$\{e\}$, $\{f,g\}$	$(0.47, 0.03)$	$H(Y_2 \mid Y_1 = 1) = 0.3274$
$\{a,b\}$	$\{a\}$, $\{b\}$	$(0.01, 0.24)$	$H(Y_3 \mid Y_2 = 0, Y_1 = 0) = 0.2423$
$\{c,d\}$	$\{c\}$, $\{d\}$	$(0.05, 0.20)$	$H(Y_3 \mid Y_2 = 1, Y_1 = 0) = 0.7219$
$\{e\}$	$\{e\}$	(0.47)	$H(Y_3 \mid Y_2 = 0, Y_1 = 1) = 0$
$\{f,g\}$	$\{f\}$,$\{g\}$	$(0.01, 0.02)$	$H(Y_3 \mid Y_2 = 1, Y_1 = 1) = 0.9183$

- partições
- Observe que $H(X) = H(Y_1, Y_2, Y_3) = 1.9323$ e lembre que

$$\begin{aligned} H(Y_1,Y_2,Y_3) &= H(Y_1) + H(Y_2 \mid Y_1) + H(Y_3 \mid Y_1, Y_2) \\ &= H(Y_1) + \sum_{i \in \{0,1\}} H(Y_2 \mid Y_1 = i) p(Y_1 = i) \\ &\quad + \sum_{i,j \in \{0,1\}} H(Y_3 \mid Y_1 = i, Y_2 = j) p(Y_1 = i, Y_2 = j) \end{aligned} \tag{457}$$

Método Guloso VII

- Árvore obtida através do método guloso

	a	b	c	d	e	f	g
p	0.01	0.24	0.05	0.20	0.47	0.01	0.02

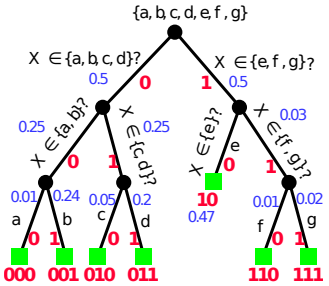


Figura 12: Árvore obtida através do método guloso (Bilmes, 2013).

- comprimento esperado do código $El = 2.53$
- entropia $H = 1.9323$
- eficiência $H/El = 0.7638$

Método Guloso VIII

- Comparando o método guloso com o método de Huffman.

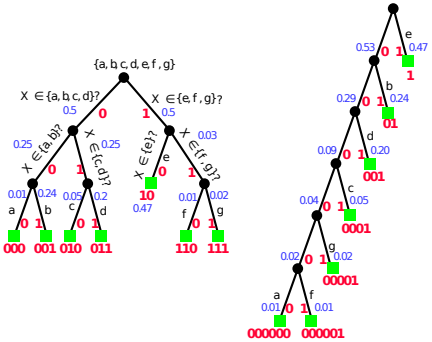


Figura 13: Guloso vs Huffman (Bilmes, 2013).

- comprimento esperado do código $El_{\text{huffman}} = 1.97$
- eficiência $H/El_{\text{huffman}} = 0.9809$
- logo, o método guloso não é ótimo

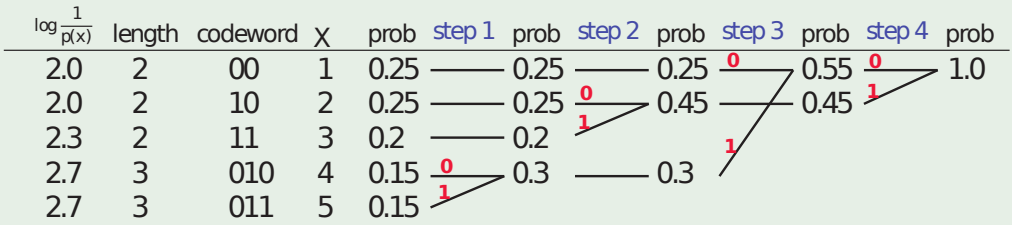
Código de Huffman I

- Procedimento
  1) Selecione os dois símbolos menos prováveis no alfabeto.
  2) Ambos terão associados as palavras mais longas e se diferirão no último bit.
  3) Combine estes dois símbolos em um símbolo auxiliar com probabilidade igual à soma das probabilidades dos dois símbolos. Adicione o símbolo auxiliar e remova os dois símbolos previamente selecionados. Repita os passos.
- Estratégia *bottom-up*.
- A estratégia é similar para $D > 2$. Neste caso poderá ser necessário utilizar símbolos fictícios no alfabeto.
- O comprimento das palavras em um código Huffman não são sempre $\leq \lceil \log 1/p_i \rceil$ (comprimento de Shannon).

Código de Huffman II

Exemplo ((Bilmes, 2013))

- $\mathcal{X} = \{1, 2, 3, 4, 5\}$ com probabilidades $(1/4, 1/4, 1/5, 3/20, 3/20)$.



- $El = 2.3$ bits e $H = 2.2855$ bits.
- Algumas palavras são maiores e outras menores que $l^*(x) = I(X) = \log 1/p(x)$.

Código de Huffman III

Exemplo

Considere $\mathcal{X}$ com probabilidades $(\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12})$.

- Entropia $H = 1.8554$ bits.
- Comprimentos de Huffman são: $L_{h1} = (2, 2, 2, 2)$ ou $L_{h2} = (1, 2, 3, 3)$.
- Comprimentos de Shannon $\lceil \log 1/p_i \rceil$ são $L_s = (2, 2, 2, 4)$ e assim $EL_s = 2.1557 > 2$. Note que $l_s(x_3) < l_{h2}(x_3)$, mas na média o comprimento esperado de Huffman é menor.

Código de Huffman é ótimo I

O código de Huffman é ótimo, i.e., $\sum_i p_i l_i$ é mínimo para comprimentos inteiros.
Para mostrar iremos fazer:

1) Lemma: alguns códigos ótimos possuem certas propriedades ($\exists$ código ótimo com estas propriedades).
2) Dado um código C_m para m símbolos, com tais propriedades, podemos criar um código mais simples satisfazendo o lemma e que será mais fácil otimizar.
3) Ao final da recurção chegaremos ao caso simples de dois símbolos em que a otimização é trivial.

Código de Huffman é ótimo II

Lema

Para toda distribuição, $\exists$ um código instantâneo ótimo (i.e., com comprimento esperado mínimo) satisfazendo simultâneamente:

1) *Se $p_j > p_k$ então $l_j \leq l_k$ (i.e., o símbolo mais provável não possui palavra com maior comprimento).*
2) *As duas maiores palavras possuem o mesmo comprimento.*
3) *As duas maiores palavras diferem apenas no último bit e correspondem aos dois símbolos menos prováveis.*

Código de Huffman é ótimo III

Demonstração.

- Suponha que C_m seja um código ótimo (então $L(C_m)$ é mínimo) e escolha j, k de forma tal que $p_j > p_k$. Precisamos mostrar que $\exists$ um código com $l_j \leq l_k$.
- Considere o código C'_m com a troca das palavras j e k, ou seja,

$$l'_j = l_k \quad \text{e} \quad l'_k = l_j \tag{458}$$

o que só pode tornar o código maior, então $L(C'_m) \geq L(C_m)$.

...

Código de Huffman é ótimo IV

Demonstração.

continuação...

- Realizando a troca, como $L(C_m)$ é mínimo, teremos

$$\begin{aligned} 0 &\leq L(C'_m) - L(C_m) = \sum_i p_i l'_i - \sum_i p_i l_i \\ &= p_j l'_j + p_k l'_k - p_j l_j - p_k l_k = p_j l_k + p_k l_j - p_j l_j - p_k l_k \\ &= p_j(l_k - l_j) - p_k(l_k - l_j) \\ &= \underbrace{(p_j - p_k)}_{>0}(l_k - l_j) \end{aligned} \tag{459}$$

- Devemos ter então $(l_k - l_j) \geq 0$, ou seja, $l_k \geq l_j$ quando $p_j > p_k$, satisfazendo assim a propriedade 1.

...

Código de Huffman é ótimo V

Demonstração.

continuação...

- Na verdade, esta propriedade é verdadeira para todos códigos ótimos.
- Propriedade 2: as palavras mais longas possuem o mesmo comprimento.
- Se as duas palavras maiores não possuem o mesmo comprimento, podemos apagar o último bit da palavra mais longa. Desta forma manteremos a propriedade de prefixo, já que a palavra mais longa é a única com o seu dado comprimento e não existe um prefixo dela que seja uma outra palavra.

irmãos após apagar

não irmãos após apagar

...

Código de Huffman é ótimo VI

Demonstração.

continuação...

- Desta forma reduzimos o comprimento esperado. Concluímos que um código ótimo deve possuir as duas palavras mais longas com o mesmo comprimento.
- Propriedade 3: as duas palavras mais longas diferem apenas no último bit e correspondem aos símbolos menos prováveis.
- Devido à propriedade 1 ($p_k < p_j \Rightarrow l_k \geq l_j$), se p_k é a menor probabilidade, então ela deve possuir associada uma palavra de comprimento que não seja menor do que qualquer outra j com $p_j > p_k$. De forma similar, se p_k é a segunda menor probabilidade, a palavra associada deve possuir comprimento que não seja menor do que qualquer outro símbolo mais provável.

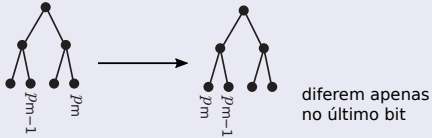
...

Código de Huffman é ótimo VII

Demonstração.

continuação...

- As duas palavras mais longas possuem o mesmo comprimento (propriedade 2) e correspondem aos dois símbolos menos prováveis.
- Se as duas maiores palavras não são irmãs, podemos trocá-las. Isto é, se $p_1 \geq p_2 \geq \ldots \geq p_m$, então fazemos a transposição ilustrada:



- Isto não altera o comprimento esperado $L = \sum_i p_i l_i$.

□

Código de Huffman é ótimo VIII

- Então, se $p_1 \geq p_2 \geq \ldots \geq p_m$, existe um código ótimo com $l_1 \leq l_2 \leq \ldots l_{m-1} = l_m$ e no qual $C(x_{m-1})$ e $C(x_m)$ diferem apenas no último bit.
- Vamos mostrar que Huffman é ótimo através da operação de criar um novo código no qual a otimização é mais simples. Este processo é repetido até que a otimização seja trivial.
- Assuma um código C_m (não necessariamente ótimo) que satisfaz as propriedades anteriores. C_m terá as seguintes palavras de código $\{w_i\}_{i=1}^{m}$.
- Huffman transforma C_m em C_{m-1} com palavras de código $\{w_i'\}_{i=1}^{m-1}$.
- Índices m e $m-1$ terão menor probabilidade e palavras maiores.

C_m	comprimento	prob. do símbolo
w_1	l_1	p_1
w_2	l_2	p_2
$\vdots$	$\vdots$	$\vdots$
w_{m-2}	l_{m-2}	p_{m-2}
w_{m-1}	l_{m-1}	p_{m-1}
w_m	l_m	p_m

Código de Huffman é ótimo IX

- Huffman realiza a seguinte operação para ir de C_m para C_{m-1}:

prob. símb.	C_{m-1}	comp. $m-1$	rel. código	rel. comp.	prob. símb.
p_1	w'_1	l'_1	$w_1 = w'_1$	$l_1 = l'_1$	p_1
p_2	w'_2	l'_2	$w_2 = w'_2$	$l_2 = l'_2$	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_{m-2}	w'_{m-2}	l'_{m-2}	$w_{m-2} = w'_{m-2}$	$l_{m-2} = l'_{m-2}$	p_{m-2}
$p_{m-1} + p_m$	w'_{m-1}	l'_{m-1}	$w_{m-1} = w'_{m-1}0$	$l_{m-1} = l'_{m-1} + 1$	p_{m-1}
			$w_m = w'_{m-1}1$	$l_m = l'_{m-1} + 1$	p_m

- w_i e l_i são as palavras e comprimento das palavras do código C_m e w'_i e l'_i do código C_{m-1}.
- As palavras e comprimentos em Huffman são definidos recursivamente. Huffman define uma relação entre palavras (e consequentemente comprimentos) ao dar um passo de um código para outro mais simples.

Código de Huffman é ótimo X

- Teremos o seguinte:

$$
\begin{aligned}
L(C_m) &= \sum_i p_i l_i \\
&= \sum_{i=1}^{m-2} p_i l'_i + p_{m-1}(l'_{m-1}+1) + p_m(l'_{m-1}+1) \\
&= \sum_{i=1}^{m-2} p_i l'_i + (p_{m-1}+p_m) l'_{m-1} + p_{m-1} + p_m \\
&= \sum_{i=1}^{m-1} p'_i l'_i + p_{m-1} + p_m \\
&= L(C_{m-1}) + \underbrace{p_{m-1}+p_m}_{\text{não involve comprimentos}}
\end{aligned} \tag{460}
$$

Código de Huffman é ótimo XI

- Desta forma, reduzimos o número de variáveis que iremos otimizar (de m para $m-1$).
- Procedimento de Huffman implica em

$$\begin{aligned} \min_{l_{1:m}} L(C_m) &= \text{const.} + \min_{l_{1:m-1}} L(C_{m-1}) = \dots \\ &= \text{const.} + \min_{l_{1:2}} L(C_2) \end{aligned} \tag{461}$$

  onde em cada passo são preservadas as propriedades.
- Ao reduzir para o caso com dois comprimentos, teremos a solução óbvia, um bit para cada símbolo, e assim podemos refazer o caminho de volta e construir o código.
- Em cada passo garantimos que estaremos obtendo a solução ótima, e assim o código criado pelo algoritmo de Huffman será ótimo.

Teorema

O procedimento de codificação de Huffman cria um código com comprimentos inteiros ótimo.

Código de Huffman é ótimo XII

- Cada símbolo terá associado uma palavra formada por um número inteiro de bits.
- Para distribuições que não são D-ádicas, poderemos utilizar até um bit extra por símbolo, na média.
- A codificação de Huffman tem a seguinte propriedade:

$$H(X) \leq L(C_{\text{Huffman}}) \leq H(X) + 1 \tag{462}$$

- A codificação de Huffman em blocos tem a propriedade:

$$H(X_{1:n}) \leq L(C_{\text{Huffman em blocos}}) \leq H(X_{1:n}) + 1 \tag{463}$$

- Podemos obter melhor resultado se codificarmos em bloco. Neste caso precisaremos calcular $p(x_{1:n})$.
- Se o alfabeto é de tamanho $|\mathcal{X}|$, precisaremos de uma tabela de tamanho $|\mathcal{X}|^n$ para armazenar todas as probabilidades.

Código de Huffman é ótimo XIII

- É difícil estimar $p(x_{1:n})$ de forma acurada. Muitas das possíveis *strings* não ocorrerão. Será necessário utilizar técnicas de suavização (*smoothing*), como por exemplo Good-Turing *Smoothing*.
- Existe também a latência introduzida pelos blocos (será necessário aguardar uma sequência de símbolos formar um bloco para poder codificar). Além disso a codificação em blocos demandará maiores recursos computacionais.
- Apesar da otimalidade do código de Huffman, duas grandes dificuldades ainda interpõem à sua aplicação: 1) o desconhecimento da real estatítica da fonte subjacente e 2) opernado em bloso, o crescimento da complexidade do algoritmo com o tamanho do bloco.

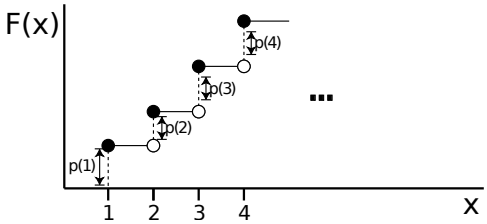
Codificação Shannon-Fano-Elias I

- Utiliza distribuição cumulativa para calcular os bits das palavras de códigos.
- Será importante para compreender a codificação aritmética.
- É necessário conhecer $p(x)$.
- $\mathcal{X} = \{1, 2, \ldots, m\}$ com $p(x) > 0$ (se houver símbolo com probabilidade nula, ele poderá ser removido).

Codificação Shannon-Fano-Elias II

Vamos definir a distribuição cumulativa

$$F(x) = \sum_{a \leq x} p(a) \tag{464}$$

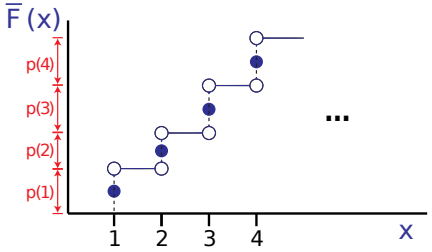


Codificação Shannon-Fano-Elias III

E definir

$$\overline{F}(x) \triangleq \sum_{a<x} p(a) + \frac{1}{2}p(x) \quad (465)$$

$$= F(x) - \frac{1}{2}p(x) \quad (466)$$



Codificação Shannon-Fano-Elias IV

- $\overline{F}(x)$ é o ponto entre $F(x-1)$ e $F(x)$, então, como $p(x) > 0$, temos

$$F(x-1) < \overline{F}(x) < F(x) \tag{467}$$

- Como $p(x) > 0$, $a \neq b \Rightarrow F(a) \neq F(b) \Leftrightarrow \overline{F}(a) \neq \overline{F}(b)$.
- Podemos utilizar $\overline{F}(a)$ como um código não singular para a (podemos utilizar a expansão binária após a vírgula, como visto anteriormente na prova de Kraft para um infinito contável de comprimentos).
- Teremos um código unicamente decodificável, porém poderemos ter palavras de tamanho infinito.
- Solução: trucar $\overline{F}(x)$ para ficar com $l(x)$ bits. A notação para tanto será $\lfloor \overline{F}(x) \rfloor_{l(x)}$. Exemplo:
  - seja $l = 4$ e $\overline{F}(x) = 0.01100100100\ldots$, então $\lfloor \overline{F}(x) \rfloor_{l(x)} = 0.0110$.
- Qual é $l(x)$ mínimo para que a decodificação unívoca seja preservada?

Codificação Shannon-Fano-Elias V

- Sabemos que

$$\overline{F}(x) - \lfloor \overline{F}(x) \rfloor_{l(x)} < \frac{1}{2^{l(x)}} \tag{468}$$

Exemplo:

- quando $l = 4$ teremos

	0.xxxx	xxxx	
$-$	0.xxxx	0000	código $\lfloor \overline{F}(x) \rfloor_4$
$=$	0.0000	xxxx	
$<$	0.0001	0000	$= 1/2^{l(x)}$

Codificação Shannon-Fano-Elias VI

- Se considerarmos $l(x) = \lceil \log 1/p(x) \rceil + 1$, então teremos

$$\frac{1}{2^{l(x)}} = \frac{1}{2} 2^{-\lceil \log 1/p(x) \rceil} \leq \frac{1}{2} 2^{-\log 1/p(x)} = \frac{p(x)}{2} \tag{469}$$

$$= \overline{F}(x) - F(x-1) \tag{470}$$

Combinando com o limite anterior, teremos

$$\overline{F}(x) - \lfloor \overline{F}(x) \rfloor_{l(x)} < \frac{1}{2^{l(x)}} < \overline{F}(x) - F(x-1) \tag{471}$$

e poderemos concluir que

$$\lfloor \overline{F}(x) \rfloor_{l(x)} > F(x-1) \tag{472}$$

- Teremos então

$$F(x-1) < \lfloor \overline{F}(x) \rfloor_{l(x)} \leq \overline{F}(x) < F(x) \tag{473}$$

Codificação Shannon-Fano-Elias VII

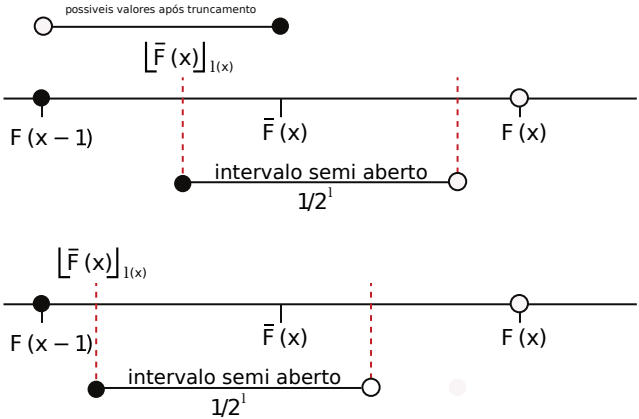
- Concluímos que $l(x) = \lceil \log 1/p(x) \rceil + 1$ bits serão suficientes para descrever x de forma não ambígua segundo a representação $\lfloor \overline{F}(x) \rfloor_{l(x)}$, uma vez que para cada x teremos $\lfloor \overline{F}(x) \rfloor_{l(x)}$ em intervalos distintos, conforme a desigualdade anterior.
- O código obtido é um código de prefixo?
- Considere a palavra $z_1 z_2 \dots z_l$ que corresponde ao intervalo semiaberto

$$[\overbrace{0.z_1 z_2 \dots z_l}^{\lfloor \overline{F}(x) \rfloor_{l(x)}}, \quad \overbrace{0.z_1 z_2 \dots z_l}^{\lfloor \overline{F}(x) \rfloor_{l(x)}} + 1/2^l) \tag{474}$$

$$= \quad [0.z_1 z_2 \dots z_l, \quad 0.z_1 z_2 \dots z_l + 0.00 \dots 1) \tag{475}$$

  que possui comprimento $1/2^l$ (este intervalo contém todos os números binários que se iniciam com $0.z_1 z_2 \dots z_l$).
- A desigualdade $F(x-1) < \lfloor \overline{F}(x) \rfloor_{l(x)} \leq \overline{F}(x) < F(x)$ e o intervalo de comprimento $1/2^l$ são representados na figura abaixo.

Codificação Shannon-Fano-Elias VIII



- Então $\lfloor \overline{F}(x) \rfloor_{l(x)} \in (F(x-1), \overline{F}(x)]$.

Codificação Shannon-Fano-Elias IX

- Como $2^{-l(x)} \leq p(x)/2 = \overline{F}(x) - F(x-1)$ e também $F(x-1) < \lfloor \overline{F}(x) \rfloor_{l(x)} \leq \overline{F}(x) < F(x)$, então teremos que os intervalos abertos são disjuntos, mesmo se $\lfloor \overline{F}(x) \rfloor_{l(x)} = \overline{F}(x)$.
- Teremos então um código de prefixo (não existirão duas palavras associadas a um mesmo intervalo e os intervalos associados a cada um são disjuntos).
- Como é suficiente ter $l(x) = \lceil \log 1/p(x) \rceil + 1$, podemos calcular o limite para o comprimento esperado.

$$L = \sum_x p(x) l(x) = \sum_x p(x)(\lceil \log 1/p(x) \rceil + 1) \leq H(X) + 2 \tag{476}$$

Codificação Shannon-Fano-Elias X

Exemplo (distribuição d-ádica)

x	$p(x)$	$F(x)$	$\overline{F}(x)$	$\overline{F}(x)$ (binário)	$l(x)$	palavra código
1	0.25	0.25	0.125	0.001	3	001
2	0.5	0.75	0.5	0.10	2	10
3	0.125	0.875	0.8125	0.1101	4	1101
4	0.125	1.0	0.9375	0.1111	4	1111

- $El = 2.75$ bits enquanto $H = 1.75$ bits.
- para o caso de Huffman teremos a árvore $(((3,4),1),2)$, e assim $El_{\text{huffman}} = 0.25 + 0.25 \times 2 + 0.125 \times 3 + 0.125 \times 3 = 1.75$.

Codificação Shannon-Fano-Elias XI

Exemplo (distribuição não d-ádica)

Notação: $0.\overline{01} = 0.010101010101\ldots$

x	$p(x)$	$F(x)$	$\overline{F}(x)$	$\overline{F}(x)$ (binário)	$l(x)$	palavra código
1	0.25	0.25	0.125	0.001	3	001
2	0.25	0.5	0.375	0.011	3	011
3	0.2	0.7	0.6	$0.1\overline{0011}$	4	1001
4	0.15	0.85	0.775	$0.110\overline{0011}$	4	1100
5	0.15	1	0.925	$0.111\overline{0110}$	4	1110

- Teremos $H = 2.285$ bits, $El = 3.5$ bits, enquanto $El_{\text{huffman}} = 2.3$ bits, sendo a árvore de Huffman $((1,(4,5)),(3,2))$.

Otimalidade Competitiva do Código de Shannon I

- Para um código em particular, podemos ter que os comprimentos para o código de Shannon são melhores do que Huffman, mas na média Huffman é melhor.
- Pergunta: Quão provável é que algum outro código unicamente decodificável seja menor do que o código de Shannon para uma palavra em particular? (para o código de Shannon é relativamente simples analisar os comprimentos, diferentemente de Huffman em que os comprimentos são definidos algoritmicamente)

Otimalidade Competitiva do Código de Shannon II

Teorema

Seja $l(x)$ o comprimento de uma palavra no código de Shannon e $l'(x)$ o comprimento de uma palavra em um outro código unicamente decodificável. Então teremos

$$\Pr\left(l(X) \geq l'(X) + c\right) \leq \frac{1}{2^{c-1}} \tag{477}$$

ou, mais formalmente expresso na forma,

$$\Pr\left(\mathbf{1}_{\{l(X) \geq l'(X)+c\}}\right) \leq \frac{1}{2^{c-1}} \tag{478}$$

Otimalidade Competitiva do Código de Shannon III

Demonstração.

$$
\begin{aligned}
\Pr\left(l(X) \geq l'(X) + c\right) &= \Pr\left(\lceil \log 1/p(X) \rceil \geq l'(X) + c\right) \\
&\quad l(x) = \lceil \log 1/p(x) \rceil \text{(código de Shannon)} \\
&\leq \Pr\left(\log 1/p(X) \geq l'(X) + c - 1\right) \\
&\quad \lceil a \rceil \leq a + 1 \\
&= \Pr\left(p(X) \leq 2^{-l'(X)-c+1}\right) \\
&= \sum_{x:p(x)\leq 2^{-l'(x)-c+1}} p(x) \qquad (479)
\end{aligned}
$$

...

Otimalidade Competitiva do Código de Shannon IV

Demonstração.

continuação...

$$\begin{aligned}
\Pr\left(l(X) \geq l'(X) + c\right) &\leq \dots \\
&= \sum_{x:p(x)\leq 2^{-l'(x)-c+1}} p(x) \\
&\leq \sum_{x:p(x)\leq 2^{-l'(x)-c+1}} 2^{-l'(x)-c+1} \\
&\leq \sum_{x} 2^{-l'(x)} 2^{-(c-1)} \leq 2^{-(c-1)} \qquad (480) \\
&\quad \text{utilizando Kraft } \sum_{x} 2^{-l'(x)} \leq 1
\end{aligned}$$

□

Otimalidade Competitiva do Código de Shannon V

- A probabilidade do código de Shannon ter comprimento esperado maior do que um outro código unicamente decodificável é exponencialmente decrescente com $c > 1$.
- Seria interessante se o código de Shannon tivesse comprimento esperado menor com maior frequência.
- Código de Shannon é ótimo para distribuições d-ádicas, já que neste caso teremos que $\log 1/p(x)$ será inteiro.

Teorema

Para uma distribuição d-ádica $p(x)$, $l(x) = \log 1/p(x)$ e $l'(x)$ é o comprimento de algum outro código de prefixo, então

$$\Pr\left(l(X) < l'(X)\right) \geq \Pr\left(l(X) > l'(X)\right) \tag{481}$$

com igualdade sse $l'(x) = l(x) \forall x$.

Otimalidade Competitiva do Código de Shannon VI

- É mais provável que o comprimento esperado de Shannon seja menor do que maior que o comprimento de outro código.

Demonstração.

Seja

$$\operatorname{sign}(t) = \begin{cases} 1, & t > 0 \\ 0, & t = 0 \\ -1, & t < 0 \end{cases} \tag{482}$$

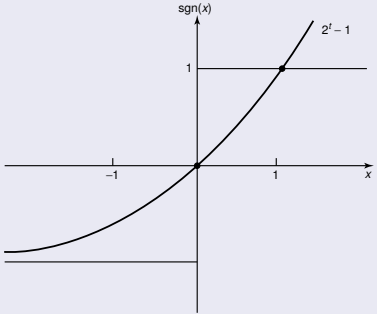
...

Otimalidade Competitiva do Código de Shannon VII

Demonstração.

continuação...

É fácil verificar pela Figura que $\text{sign}(t) \leq 2^t - 1$ para $t = 0, \pm 1, \pm 2, \ldots$.



...

Otimalidade Competitiva do Código de Shannon VIII

Demonstração.

continuação...

Teremos então

$$
\begin{aligned}
&\Pr\left(l'(X) < l(X)\right) \\
-\Pr\left(l'(X) > l(X)\right) &= \sum_{x:l'(x)\leq l(x)} p(x) - \sum_{x:l'(x)>l(x)} p(x) \\
&\quad \text{como } \{l'(x) \leq l(x)\} \cap \{l'(x) > l(x)\} = \emptyset \\
&= \sum_{x} p(x) \operatorname{sign}\left(l(x) - l'(x)\right) \\
&= E\left[\operatorname{sign}\left(l(X) - l'(X)\right)\right] \\
&\leq \sum_{x} p(x) \left(2^{l(x)-l'(x)} - 1\right) \qquad (483)
\end{aligned}
$$

...

Otimalidade Competitiva do Código de Shannon IX

Demonstração.

continuação...

Como $p(x)$ é d-ádica, $p(x) = 2^{-l(x)}$, teremos

$$\begin{aligned} \Pr\left(l'(X) < l(X)\right) & \\ -\Pr\left(l'(X) > l(X)\right) &\leq \sum_x p(x)\left(2^{l(x)-l'(x)} - 1\right) \\ &= \sum_x 2^{-l(x)}\left(2^{l(x)-l'(x)} - 1\right) \\ &= \sum_x 2^{-l'(x)} - \sum_x 2^{-l(x)} \\ &= \sum_x 2^{-l'(x)} - 1 \end{aligned} \tag{484}$$

...

Otimalidade Competitiva do Código de Shannon X

Demonstração.

continuação...

$$\begin{aligned} \Pr\left(l'(X) < l(X)\right) & \\ -\Pr\left(l'(X) > l(X)\right) &\leq \dots \\ &= \sum_x 2^{-l'(x)} - 1 \\ &\leq 1 - 1, \text{ já que } l'(x) \text{ satisfaz Kraft} \\ &= 0 \end{aligned} \tag{485}$$

Assim., mostramos que $\Pr\left(l'(X) < l(X)\right) \leq \Pr\left(l'(X) > l(X)\right)$, como desejado.

□

Jogos de Shannon I

- Redundância está presente em todos lugares. Em um língua existe redundância no nível de sentenças, no nível de palavras e no nível de caracteres.
- Exemplos:

  "Não se ama duas vezes a mesma ".

  "A gratidão de quem recebe um benefício é bem menor que o prazer daquele de quem o ".

  (Machado de Assis)

- Shannon percebeu isto e propôs uma maneira de estimar a entropia.
- Assuma o alfabeto 'a' a 'z' mais espaço (27 caracteres).
- Um caractere é dado a cata instante e uma pessoa deve adivinhar qual é o próximo.

  T H E R E - I S - NO- R E V E R S E - ON- A - MO T O R C Y C L E -
  1 1 1 5 1 1 2 1 1 2 1 1 15 1 17 1 1 1 2 1 3 2 1 2 2 7 1 1 1 1 4 1 1 1 1 1

Jogos de Shannon II

- O número de tentativas para adivinhar a letra correta pode ser vista como um 'código' para a *string*. Teremos assim um mapeamento de letras em inteiros:

$$C : \{'a', 'b', 'c', \ldots, 'z', ' '\} \to \{1, 2, \ldots, 27\} \tag{486}$$

- Usualmente o número de tentativas será pequeno. Aquilo que é mais dedutível terá maior probabilidade e menor número de tentativas serão necessárias.
- Seja g_t o número de tentativas até acertar na posição t, então $\log g_t$ será o número de bits necessários para representar g_t.
- A taxa de entropia deste processo é

$$\hat{H}(X) \approx \frac{1}{n} \sum_{t=1}^{n} \log g_t. \tag{487}$$

Jogos de Shannon III

- Supondo que $x_1, x_2, \ldots$ é um processo estocástico com taxa de entropia da forma $H(X_t \mid X_{t-1}, \ldots, X_1)$, então $p(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1)$ é a probabilidade da letra x_t no instante t. Teremos então

$$g_t \approx \frac{1}{p(x_t \mid x_{t-1}, x_{t-2}, \ldots, x_1)}. \tag{488}$$

- Ao realizar tal codificação teremos que os números pequenos serão muito mais frequentes do que os grandes. Para o exemplo anterior, 'There is no reverse on a motorcycle', será associada a seguinte sequência de símbolos: 1, 1, 1, 5, 1, 1, 2, 1, 1, 2, 1, 1, 15, 1, 17, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 7, 1, 1, 1, 1, 4, 1, 1, 1, 1, 1.
- Devemos ter um bom resultado na compressão desta sequência.
- Para realizar a decodificação precisaremos de um gêmeo idêntico, que faça as mesmas adivinhações, ou seja, realizará g_t tentativas até acertar.

Jogos de Shannon IV

- Outra alternativa seria criar uma longa tabela com todos históricos possíveis e memorizar o esquema as tentativas que um humano faria. Para uma *string* de tamanho L temos 27^L combinações, tornando esta abordagem impraticável.
- Uma alternativa viável seria considerar $P(X_t \mid X_{t-1}, \ldots, X_1) \approx P(X_t \mid T(X_1, \ldots, X_{t-1}))$, onde $T(X_1, \ldots, X_{t-1})$ é uma estatística sobre as $t-1$ amostras passadas e a cada instante basta ajustar a estatística.
- Este esquema introduzido por Shannon nos anos 1950 é a base para a codificação aritmética.

Codificação Aritmética I

- Utilizada em compressão de documentos: DjVU, PDF, JPEG.
- Assumir o seguinte modelo probabilístico da fonte:

$$p(x_{1:n}) = \prod_{i=1}^{n} p(x_i) \quad \text{i.i.d.} \tag{489}$$

ou, de forma alternativa,

$$p(x_{1:n}) = p(x_1) \prod_{i=2}^{n} p(x_i \mid x_{i-1}) \quad \text{modelos de Markov de 1a ordem.} \tag{490}$$

- A cada símbolo, utilizamos a probabilidade condicional para encontrar a probabilidade do próximo símbolo. É possível lidar de forma simples com modelos complexos adaptativos da fonte.

Codificação Aritmética II

- Exemplo: $\mathcal{X} = a, e, i, o, u, !$, então $|\mathcal{X}| = 6$, no qual acrescentamos um símbolo extra !, o símbolo de termino.
- A sequência produzida pela fonte $X_1, X_2, \ldots$ não precisa ser i.i.d.
- Vamos assumir que $p(x_n \mid x_1, x_2, \ldots, x_{n-1})$ é dado ao codificador e decodificador. Teremos uma descrição algorítmica desta distribuição, que será uma descrição finita.
- Assim como na codificação de Shannon-Fano-Elias, iremos dividir o intervalo unitário em segmentos de acordo com as probabilidades $p(X_1 = x)$ para $x \in \{a, e, i, o, u, !\}$.

Codificação Aritmética III

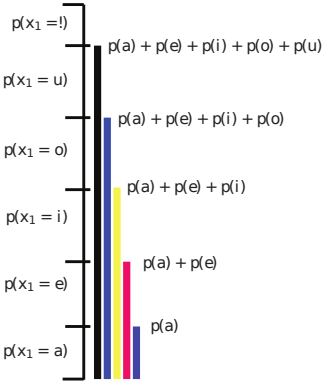


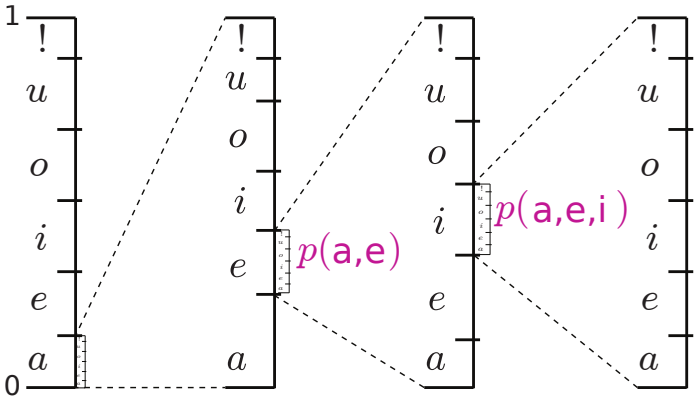
Figura 14: Codificação aritmética - exemplo (Bilmes, 2013).

- Cada subintervalo pode ser subdividido em segmentos de comprimento relativo $p(X_2 = x_2|X_1 = x_1)$ ou comprimento efetivo $p(X_2 = x_2, X_1 = x_1)$.

Codificação Aritmética IV

- Os comprimentos relativos podem ser maiores, menores ou iguais, ou seja, $p(X_2 = x_2) \gtreqless p(X_2 = x_2|X_1 = x_1)$.
- A Figura abaixo ilustra as probabilidades considerando que a seguinte sequência ocorreu $x_1 = a, x_2 = e, x_3 = i$ (Bilmes, 2013).

Codificação Aritmética V



- O comprimento absoluto para a sequência 'ae' é $p(X_1 = a, X_2 = e) = p(X_1 = a)p(X_2 = e \mid X_1 = a)$.
- Os intervalos tornam-se exponencialmente menores com n.

Codificação Aritmética VI

- A cada passo, os comprimentos relativos podem mudar dependendo do passado. No instante $t = 1$, o comprimento relativo para 'a' era $p(a)$; em $t = 2$, o comprimento relativo para 'a' será $p(a \mid X_1)$, que pode mudar dependendo do valor que X_1 assumiu.
- O número de bits necessários para especificar um determinado sub-intervalo é aproximadamente o mesmo que anteriormente, ou seja, o número de bits que serão acrescentados em cada etapa é aproximadamente o mesmo. Será o mesmo se os comprimentos relativos não mudarem, ou seja, se a probabilidade condicional não mudar.
- Se um símbolo fica muito provável, o comprimento relativo do intervalo é grande e assim utilizará poucos bits; por outro lado, se um símbolo torna-se pouco provável, o comprimento relativo do intervalo diminui e serão necessários mais bits para distinguir este símbolo. (Exemplo: o intervalo correspondente a $[0.0110, 0.0111)$ é menor do que o intervalo correspondente a $[0.10, 0.11)$.)

Codificação Aritmética VII

- Procedimento para codificar.
- Para o i-ésimo símbolo X_i, considere
  - o limite inferior do intervalo

$$L_n(i \mid x_1, x_2, \ldots, x_{n-1}) = \sum_{j=1}^{i-1} p(x_n = j \mid x_1, x_2, \ldots, x_{n-1}) \tag{491}$$

  - o limite superior do intervalo

$$U_n(i \mid x_1, x_2, \ldots, x_{n-1}) = \sum_{j=1}^{i} p(x_n = j \mid x_1, x_2, \ldots, x_{n-1}) \tag{492}$$

  - Note que $U_n = L_n + p(x_n = i \mid x_1, x_2, \ldots, x_{n-1})$, ou seja, teremos $L_n \leq U_n$ com igualdade sse $p(x_n = i \mid x_1, x_2, \ldots, x_{n-1}) = 0$.
- Para codificar o n-ésimo símbolo, iremos dividir o $(n-1)$-ésimo intervalo definido pelo intervalo semiaberto $[L_n, U_n)$.

Codificação Aritmética VIII

Para o exemplo dado, considere o intervalo inicial $[0, 1)$. Este será dividido de acordo com as probabilidades dos símbolos.

$$a \leftrightarrow [L_1(a), U_1(a)) = [0, p(X_1 = a)) \tag{493}$$

$$e \leftrightarrow [L_1(e), U_1(e)) = [p(X_1 = a), p(X_1 = a) + p(X_1 = e)) \tag{494}$$

$$i \leftrightarrow [L_1(i), U_1(i)) = [p(a) + p(e), p(a) + p(e) + p(i)) \tag{495}$$

$$o \leftrightarrow [L_1(o), U_1(o)) = [p(a) + p(e) + p(i), p(a) + p(e) + p(i) + p(o)) \tag{496}$$

$$u \leftrightarrow [L_1(u), U_1(u)) = [\sum_{x \in \{a,e,i,o\}} p(x), \sum_{x \in \{a,e,i,o,u\}} p(x)) \tag{497}$$

$$! \leftrightarrow [L_1(!), U_1(!)) = [\sum_{x \in \{a,e,i,o,u\}} p(x), 1) \tag{498}$$

Codificação Aritmética IX

- Vamos utilizar o algoritmo abaixo para codificar uma *string* $x_1, x_2, \ldots, x_N$.
- Para tanto iremos determinar os intervalos $[l, u)$ em cada passo t, onde l é o limite inferior e u o limite superior.

$l \leftarrow 0$
$u \leftarrow 1$
$p \leftarrow u - l$ ▷ neste caso será 1
for n = 1 ... N **do** ▷ calcule $\forall i \in \mathcal{X}$, U_n e L_n como dado acima
  $u \leftarrow l + pU_n(x_n \mid x_1, \ldots, x_{n-1})$
  $l \leftarrow l + pL_n(x_n \mid x_1, \ldots, x_{n-1})$
  $p \leftarrow u - l$
end for

- Ao final de N iterações, teremos um intervalo final $[l, u)$. Para codificar, basta transmitir uma *string* binária qualquer de um número neste intervalo.

Codificação Aritmética X

- Por outro lado, é possível gerar a sequência binária em tempo de execução, para tanto iremos começar a escrever os bits à medida que sabemos, de forma não ambígua, que estamos em um determinado intervalo.
- De forma análoga à codificação de Shannon-Fano-Elias, se o intervalo atual é $[0.100101, 0.100110)$, podemos enviar os bits que constituem o prefixo comum 1001, já que isto não será alterado, independente do que ocorrer depois.

Codificação Aritmética XI

Exemplo ((MacKay, 2003))

Suponha $x \in \{a, b, \square\}$, onde $\square$ é o símbolo de término.

- Vamos considerar a codificação da *string* $bbba\square$.

-	$p(a) = 0.425$	$p(b) = 0.425$	$p(\square) = 0.15$
b	$p(a\|b) = 0.28$	$p(b\|b) = 0.57$	$p(\square\|b) = 0.15$
bb	$p(a\|bb) = 0.21$	$p(b\|bb) = 0.64$	$p(\square\|bb) = 0.15$
bbb	$p(a\|bbb) = 0.17$	$p(b\|bbb) = 0.68$	$p(\square\|bbb) = 0.15$
$bbba$	$p(a\|bbba) = 0.28$	$p(b\|bbba) = 0.57$	$p(\square\|bbba) = 0.15$

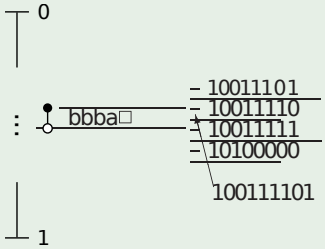
...

Codificação Aritmética XII

Exemplo ((MacKay, 2003))

continuação...

- A figura ilustra os intervalos que teremos ao final da sequência $bbba\square$.



...

Codificação Aritmética XIII

Exemplo ((MacKay, 2003))

continuação...

- Temos subintervalos semiabertos em $[0, 1)$.
- Os intervalos $[l, u)$ são dados pela probabilidade condicional $p(x_i|x_1, x_2, \ldots, x_{i-1})$.
- Temos a representação de um prefixo $b_1 b_2 \ldots b_k$ por um intervalo da forma $[0.b_1 b_2 \ldots b_k, 0.b_1 b_2 \ldots b_k + 2^{-k})$ para $b_i \in \{0, 1\}$.
- A palavra gerada pela codificação será 100111101, que determina um intervalo dentro do intervalo final.

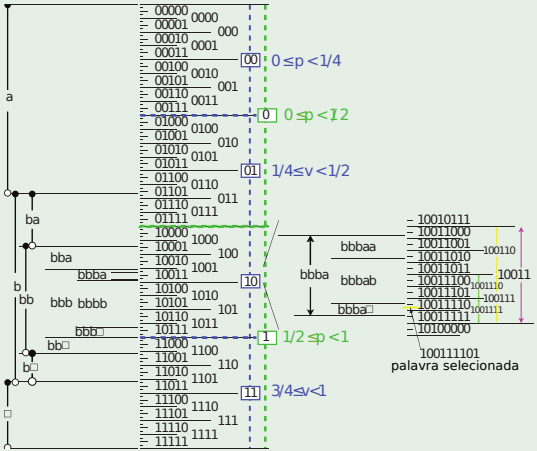
...

Codificação Aritmética XIV

Exemplo ((MacKay, 2003))

Codificação Aritmética XV

continuação...



Codificação Aritmética XVI

- Decodificação.
- Para decodificar uma *string* $\alpha = 0.z_1 z_2 z_3 \ldots$ utilizaremos o seguinte algoritmo:

$l \leftarrow 0$
$u \leftarrow 1$
$p \leftarrow u - l$
while não receber o simbolo final $\square$ **do**
    encontre i tal que $L_n(i|x_1, \ldots, x_{n-1}) \leq \frac{\alpha - l}{u - l} < U_n(i|x_1, \ldots, x_{n-1})$
    $u \leftarrow l + pU_n(i \mid x_1, \ldots, x_{n-1})$
    $l \leftarrow l + pL_n(i \mid x_1, \ldots, x_{n-1})$
    $p \leftarrow u - l$
end while

Codificação Aritmética XVII

- Determinação do número de bits.
- Um determinado número no intervalo final $[L_n, U_n)$ pode ser arbitrariamente longo. Precisaremos enviar apenas o suficiente para identificar a *string* original de forma univoca.
- Defina

$$F_n(i \mid x_1, x_2, \ldots, x_{n-1}) = \frac{1}{2}[L_n(i) + U_n(i)], \tag{499}$$

  e $\lfloor F_n(i \mid x_1, x_2, \ldots, x_{n-1}) \rfloor_l$ é F_n truncado com l bits.
- Poderíamos definir a número de bits utilizado em cada estágio

$$l(x_n|x_1, \ldots, x_{n-1}) = \lceil \log 1/p(x_n|x_1, \ldots, x_{n-1}) \rceil + 1. \tag{500}$$

Codificação Aritmética XVIII

- Ao invés disso, utilizaremos o comprimento de Shannon para o código inteiro

$$l(x_{1:n}) = \lceil \log 1/p(x_{1:n}) \rceil + 1. \tag{501}$$

À medida que os símbolos chegam ao codificador, é possível calcular

$$p(x_{1:n}) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) \tag{502}$$

Assim, a cada passo, saberemos quantos bits a mais são necessários em cada passo.

- Pelo mesmo argumento dado no código de Shannon-Fano-Elias, teremos um código de prefixo e desta forma, unicamente decodificável.

Codificação Aritmética XIX

- O comprimento esperado será dado por

$$\begin{aligned} El(x_{1:n}) &= \sum_{x_{1:n}} p(x_{1:n}) l(x_{1:n}) \\ &= \sum_{x_{1:n}} p(x_{1:n}) \left(\lceil \log 1/p(x_{1:n}) \rceil + 1 \right) \\ &\leq -\sum_{x_{1:n}} p(x_{1:n}) (\log p(x_{1:n}) + 2) \\ &= H(x_{1:n}) + 2 \end{aligned} \tag{503}$$

- Por símbolo teremos $El \leq H(x_{1:n})/n + 2/n \to H(\mathcal{X})$.
- Temos um *stream code* $\neq$ *block code*.
- Ainda temos o problema de estimar $p(x_n | x_1, \ldots, x_{n-1})$.
- Outro problema é que os valores das probabilidades ficam exponencialmente pequenos, podendo ocorrer erro de precisão numérica. É necessário então expandir os intervalos de tempos em tempos.

Estimar $p(x_n|x_1, \ldots, x_{n-1})$ I

- Ainda temos o problema de estimar $p(x_n|x_1, \ldots, x_{n-1})$.
- Vamos utilizar um modelo adaptativo.
  - Modelo de Dirichlet

$$p(a \mid x_{1:n-1}) = \frac{N(a|x_{1:n-1}) + \alpha}{\sum_{a'} \left(N(a' \mid x_{1:n-1}) + \alpha\right)} \tag{504}$$

    onde $\alpha \geq 0$.
- Temos um problema de estimação de densidade.

Regra de Laplace: derivação Bayesiana I

- Vamos assumir um alfabeto binário $\mathcal{X} = \{0, 1\}$.
- O número de ocorrências de 0 e 1 são dados respectivamente por

$$N_0 = N(0|x_{1:n}) \quad \text{e} \quad N_1 = N(1|x_{1:n}), \tag{505}$$

  sendo que $n = N_0 + N_1$.
- Vamos assumir p_0, p_1 e $p_\square$ as probabilidades para ps símbolos 0, 1 e o símbolo terminal $\square$.
- O comprimento de uma *string* possui distribuição geométrica

$$p(l) = (1 - p_\square)^l p_\square. \tag{506}$$

  Desta forma, $\mathrm{E}l = 1/p_\square$ e $\mathrm{Var}(l) = (1 - p_\square)/p_\square^2$.
- As realizações das v.a.s na sequência $X_{1:N}$ são i.i.d., desta forma teremos

$$p(x_{1:N} \mid p_0, N) = p_0^{N_0} p_1^{N_1}. \tag{507}$$

  Esta é a função de verossimilhança dos parâmetros dados os dados.

Regra de Laplace: derivação Bayesiana II

- Vamos utilizar que *a priori* temos distribuição uniforme, i.e., $\Pr(p_0) = 1$ para $p_0 \in [0, 1]$. Note que isto é feito para descrever a incerteza com relação à real distribuição que nos é desconhecida. A distribuição p não é aleatória, mas incerta. Atribuímos uma distribuição a p para expressar a incerteza, não para atribuir aleatoriedade a p; mas, matematicamente, será tratado da mesma forma, como se a distribuição p fosse aleatória.
- Teremos assim

$$\begin{aligned} \Pr(p_0 \mid x_{1:n}, N) &= \frac{\Pr(x_{1:n} \mid p_0, N) \Pr(p_0)}{\Pr(x_{1:n} \mid N)} \\ &= \frac{p_0^{N_0} p_1^{N_1}}{\Pr(x_{1:n} \mid N)} \end{aligned} \tag{508}$$

Regra de Laplace: derivação Bayesiana III

- Vamos utilizar a integral Beta dada a seguir

$$B(a,b) \equiv \int x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{509}$$

onde $\Gamma(\cdot)$ é a função gamma, uma extensão da função fatorial, com argumento deslocado de 1. Para n inteiro, teremos $\Gamma(n) = (n-1)!$. A função gamma é definida para todos números complexos, exceto os inteiros não positivos, pela seguinte integral impropria convergente:

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx. \tag{510}$$

Regra de Laplace: derivação Bayesiana IV

- Utilizando a função Beta, teremos

$$\begin{aligned} \Pr(x_{1:n} \mid N) &= \int_0^1 \Pr(x_{1:n}, p_0 \mid N) dp_0 = \int_0^1 p_0^{N_0} p_1^{N_1} \Pr(p_0) dp_0 \\ &= \frac{\Gamma(N_0+1)\Gamma(N_1+1)}{\Gamma(N_0+N_1+2)} = \frac{N_0! N_1!}{(N_0+N_1+1)!} \end{aligned} \tag{511}$$

Regra de Laplace: derivação Bayesiana V

- Para realizar a predição, queremos saber qual é a probabilidade do próximo símbolo ser igual a 0, por exemplo, dado que temos uma sequência de N observações, ou seja,

$$\begin{aligned}
\Pr(X_{n+1}=0 \mid x_{1:n}, N) &= \int_0^1 \underbrace{\Pr(0 \mid p_0)}_{=p_0} \Pr(p_0 \mid x_{1:n}, N) dp_0 \\
&= \int_0^1 p_0 \frac{p_0^{N_0} p_1^{N_1}}{\Pr(x_{1:n} \mid N)} dp_0 \\
&= \int_0^1 \frac{p_0^{N_0+1} p_1^{N_1}}{\Pr(x_{1:n} \mid N)} dp_0 \\
&= \left(\frac{(N_0+1)!N_1!}{(N_0+N_1+2)!} \right) / \left(\frac{N_0!N_1!}{(N_0+N_1+1)!} \right) \\
&= \frac{N_0+1}{N_0+N_1+2} \quad \text{regra de Laplace} \qquad (512)
\end{aligned}$$

Regra de Laplace: derivação Bayesiana VI

- Para a abordagem de Dirichlet, teremos a seguinte aproximação:

$$Pr(X_{n+1} = 0 \mid x_{1:n}, N) = \frac{N_0 + \alpha}{N_0 + N_1 + 2\alpha} \tag{513}$$

- Para calcular esta probabilidade ao longo do processo de codificação, não será necessário armazenar toda a história, mas apenas alguma estatística sobre os dados, como por exemplo, o número de ocorrências de cada símbolo.

Regra de Laplace: derivação Bayesiana VII

Regra da Sucessão

Regra da Sucessão é a formulação introduzida no Século XVIII por Pierre-Simon Laplace para lidar com o 'problema do nascer do sol'.
Se repetirmos um experimento, que sabemos previamente que pode resultar em sucesso ou falha, n vezes de forma independente e observamos s sucessos, então como podemos estimar a probabilidade de que a próxima realização será um sucesso? (Qual é a probabilidade de que o sol nascerá amanhã, visto que nos últimos 5 mil anos, ou 1826251 dias, foi observado que o sol nasceu todas as vezes?) ...

Regra de Laplace: derivação Bayesiana VIII

Regra da Sucessão

continuação...

Ou seja, se $X_1, \ldots, X_{n+1}$ são v.a.s condicionalmente independentes, que assumem valor 0 ou 1, então, se não sabemos nada além disso, teremos

$$P(X_{n+1} = 1 \mid X_1 + \ldots + X_n = s) = \frac{s+1}{n+2}. \tag{514}$$

Note que, se não soubéssemos que ambos, sucesso e fracasso, são possíveis, então teríamos a seguinte probabilidade

$$P'(X_{n+1} = 1 \mid X_1 + \ldots + X_n = s) = \frac{s}{n}. \tag{515}$$

...

Regra de Laplace: derivação Bayesiana IX

Regra da Sucessão

continuação...

Seja $X_i = 1$ se observamos um sucesso na i-ésima realização da v.a. e 0 caso contrário, com probabilidade p de sucesso, teremos então uma distribuição de Bernoulli. Vamos supor que $X_1, \ldots, X_n$ são independentes, dado p. Pelo teorema de Bayes, para encontrar a distribuição de probabilidade condicional de p dado X_i, $i = 1, \ldots, n$, devemos multiplicar a medida da probabilidade a priori dada a p pela função de verosimilhança

$$L(p) = P(X_1 = x_1, \ldots, X_n = x_n \mid p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^s(1-p)^{n-s} \tag{516}$$

onde $s = x_1 + \ldots + x_n$ é o número de sucessos e n o número de tentativas. ...

Regra de Laplace: derivação Bayesiana X

Regra da Sucessão

continuação...

A função densidade de probabilidade a priori que expressa ignorância total sobre p (exceto pela fato de que sabemos que não é 1, nem 0, ou seja, existem duas possibilidades de fato: sucesso ou falha) é uniforme, sendo igual a 1 para $0 < p < 1$ e 0 (zero) caso contrário. Para obter a constante de normalização faremos

$$\int_0^1 p^s (1-p)^{n-s} dp = \frac{s!(n-s)!}{(n+1)!} \tag{517}$$

...

Regra de Laplace: derivação Bayesiana XI

Regra da Sucessão

continuação...

A função densidade de probabilidade a posteriori é dada pelo produto da função densidade de probabilidade a priori com a função de verossimilhança e normalizando. Teremos assim

$$f(p) = \frac{(n+1)!}{s!(n-s)!} p^s (1-p)^{n-s}. \tag{518}$$

Esta é a distribuição beta com valor esperado

$$\int_0^1 p f(p) dp = \frac{s+1}{n+2}. \tag{519}$$

...

Regra de Laplace: derivação Bayesiana XII

Regra da Sucessão

continuação...

Como a probabilidade condicional de sucesso na próxima realização, dado o valor de p, é apenas p, a lei da probabilidade total diz que a probabilidade de sucesso na próxima realização é apenas o valor esperado de p. Como tudo isto é condicional aos dados observados X_i, para $i = 1, \ldots, n$, temos

$$P(X_{n+1} = 1 \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{s+1}{n+2}. \tag{520}$$

Complexidade de Kolmogorov I

- A complexidade de Kolmogorov de uma *string* x, denotada por $K(x)$, é o comprimento do menor programa capaz de gerar x, quando executado em um computador universal $\mathcal{U}$.
- Esta definição é algorítmica.
- Uma *string*, cujo menor programa é do mesmo tamanho que a própria *string*, é dita algoriticamente aleatória (ou algoritmicamente incompressível).
- exemplo
  $x =$ 'abababababababababababababababab' pode ser descrito como 'ab 16 vezes'.
  $x =$ '4c1j5b2p0cv4w1x8rx2y39umgw5q85s7' aparentemente não possui uma descrição simples (talvez a descrição mais simples de x seja o próprio x).
- $K(x)$ é algorítmico e às vezes relacionado com a entropia, mas não existe uma maneira prática de computá-lo.
- Existe alguma maneira puramente algorítmica de comprimir (exceto K) e que atinja a taxa de entropia no limite?

Complexidade de Kolmogorov II

- Sendo um método algorítmico, desejamos que ele não necessite de calcular a distribuição probabilística governando os símbolos.
- Queremos um compressor universal que alcance o limite da taxa de entropia.
- Ideia básica do Lempel-Ziv é memorizar as *strings* que já ocorreram, sem precisar lidar com a distribuição da fonte (universal) e ainda conseguir atingir o limite da taxa de entropia.
- O algoritmo de Lempel-Ziv é utilizado no `gzip`, amplamente utilizado na compressão de textos.

Compressão Lempel Ziv I

- A sequência produzida pela fonte é analisada (*parsed*) nas menores frases que ainda não ocorreram até então (histórico).
- exemplo: a frase 'a␣casa␣caiu' terá a seguinte análise:
  a, ␣, c, as, a␣, ca, i, u
- exemplo binário: a *string* '1011010100010 ...' será analisada com
  1, 0, 11, 01, 010, 00, 10, ...
- Para codificar, fornecemos o local do prefixo (*string* que já ocorreu antes, exceto o símbolo final) e então adiciona o índice do símbolo final. Utiliza-se 0 como ponteiro nulo, indicando que a *string* não ocorreu antes.
- exemplo: 'a␣casa␣caiu'

frase	a	␣	c	as	a␣	ca	i	u
posição	1	2	3	4	5	6	7	8
código	(0,a)	(0,␣)	(0,c)	(1,s)	(1,␣)	(3,a)	(0,i)	(0,u)

Compressão Lempel Ziv II

- exemplo binário: '1011010100010 ...'

frase	1	0	11	01	010	00	10
posição	1	2	3	4	5	6	7
código	(0,1)	(0,0)	(1,1)	(2,1)	(4,0)	(2,0)	(1,0)

- De forma geral, à medida que a codificação se procede, os ponteiros referenciam *strings* cada vez mais longas, fazendo assim com que a taxa de compressão melhore (assumindo que existe regularidade na fonte, gerando *strings* repetidas).

Lempel Ziv - Codificação Binária I

Vamos analisar a codificação binária.

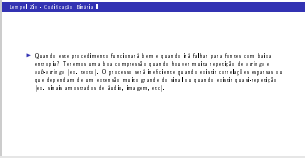
- O número de bits necessários para a codificação dependerá do número de frases encontradas na *string* gerada pela fonte.
- $c(n)$ é o número de frases de uma *string* de comprimento n quando analisada seguindo o algoritmo descrito anteriormente.
- Serão necessários $\lceil \log c(n) \rceil + 1$ bits, onde o último bit ($+1$) é adicionado para descrever o bit final.
- Para o exemplo dado anteriormente ('1011010100010'), teremos $c(n) = 7$, e assim:

frase	1	0	11	01	010	00	10
posição	1	2	3	4	5	6	7
código	(0,1)	(0,0)	(1,1)	(2,1)	(4,0)	(2,0)	(1,0)
código	(000,1)	(000,0)	(001,1)	(010,1)	(100,0)	(010,0)	(001,0)

- O processo de decodificação é simples: basta criar uma lista das *strings* que já foram encontradas e, quando encontrar (i, j), escreva na saída a *string* armazenada na posição i e em seguida escreva j.

Lempel Ziv - Codificação Binária II

- Quando este procedimento funcionará bem e quando irá falhar para fontes com baixa entropia? Teremos uma boa compressão quando houver muita repetição de *strings* e *sub-strings* (ex. texto). O processo será ineficiente quando existir correlações esparsas ou que dependam de um extensão muito grande do sinal ou quando existir quasi-repetição (ex. sinais amostrados de áudio, imagem, etc).



(merriam-webster)
Definition of **parse**:
transitive verb

1. to divide (a sentence) into grammatical parts and identify the parts and their relations to each other
2. to examine in a minute way :analyze critically

intransitive verb

1. to give a grammatical description of a word or a group of words
2. to admit of being parsed

Lempel Ziv I

- Vamos considerar um alfabeto binário $\mathcal{X} = \{0, 1\}$.

Definição (Análise (*parsing*))

Uma análise (*parsing*) S de uma *string* $x_1, x_2, \ldots, x_n$ é uma divisão da *string* em frases, separadas por vírgulas (ou outro separador qualquer).

Definição (Análise Distinta (*distinct parsing*))

Uma análise distinta (*distinct parsing*) é uma análise (*parsing*) em que não existe duas frases idênticas.

- Ex. 01101101. Uma análise possível é: 0,11,0,11,01. Mas esta não é distinta.
- Uma análise distinta de 01101101 seria 0,1,10,11,01.
- Lempel-Ziv produz uma análise distinta da sequência da fonte.
- Seja $c(n)$ o número de frases na análise LZ de uma *string* de comprimento n. Então $c(n)$ depende da *string* $x_{1:n}$. $(c = c(n) = c(x_{1:n}))$

Lempel Ziv II

- Após a compressão, teremos uma sequência de $c(n)$ pares da forma (ponteiro, bit), onde o ponteiro requer $\lceil \log c(n) \rceil$ bits.
- O comprimento da sequência comprimida será

$$c(n)(\lceil \log c(n) \rceil + 1) \text{ bits} \tag{521}$$

- Queremos que a compressão LZ atinja o limite da taxa de entropia, ou seja, queremos

$$\frac{c(n)(\lceil \log c(n) \rceil + 1)}{n} \to H(X) \tag{522}$$

para uma sequência estacionária ergódica $x_{1:n}$, ou seja,

$$\frac{1}{n} \mathrm{E}_{p(x_{1:n})} \left[c(X_{1:n})(\lceil \log c(X_{1:n}) \rceil + 1) \right] = H(X) \tag{523}$$

Dominância I

- Dominância (*little-o notation*)

$$o(g(n)) \triangleq \{f(n) : \forall c > 0, \exists n_0 > 0 \quad / \quad 0 \leq f(n) \leq cg(n), \forall n \geq n_0\} \tag{524}$$

  ou seja, $o(g(n))$ é o conjunto de todas as funções dominadas por g.
- Intuitivamente, isto significa que $g(n)$ cresce muito mais rápido do que uma função $f(n) \in o(g(n))$ qualquer (o crescimento de $f(n)$ não é nada comparado ao de $g(n)$).
- Se $g(n)$ não é nulo, então para qualquer $f \in o(g(n))$, teremos $\lim_{n\to\infty} f(n)/g(n) = 0$.
- As funções $o(1)$ são funções que tendem a zero no limite, isto é

$$o(1) \triangleq \{f(n) : \forall c > 0, \exists n_0 > 0 \quad / \quad 0 \leq f(n) < c, \forall n \geq n_0\} \tag{525}$$

  isto é, se $f \in o(1)$, então $\lim_{n\to\infty} f(n) = 0$.

Lempel-Ziv: Demonstração I

Lema (limite superior do número de frases)

O número de frases $c(n)$ em qualquer análise (parsing) distinta de sequências binárias $x_{1:n}$ satisfaz:

$$c(n) \leq \frac{n}{(1-\epsilon_n)\log n} \tag{526}$$

onde $\epsilon \to 0$ e $n \to \infty$. Ou seja,

$$c(n) \leq \frac{n}{\log n}(1+o(1)) \tag{527}$$

nota:

$$\frac{1}{1-\epsilon} = \underbrace{\frac{1}{1-\epsilon} - 1}_{\in o(1)} + 1 \tag{528}$$

Lempel-Ziv: Demonstração II

Demonstração.

- Na demonstração utilizaremos a seguinte igualdade:

$$\sum_{l=1}^{k} 2^l = \sum_{l=0}^{k} 2^l - 1 = 2^{k+1} - 2 \tag{529}$$

- Seja n_k a soma de todos os comprimentos de todas as *strings* distintas de comprimento $\leq k$ (a soma do comprimento de todas as *strings* no dicionário, ou o tamanho do dicionário). Para um dado comprimento j, existem 2^j *strings* binárias distintas.

$$n_k = \sum_{j=1}^{k} j 2^j \tag{530}$$

...

Lempel-Ziv: Demonstração III

Demonstração.

continuação...

Na demonstração abaixo iremos utilizar que $\sum_{l=1}^{k} 2^l = 2^{k+1} - 2$.

$$
\begin{aligned}
n_k &= \sum_{j=1}^{k} j2^j = \sum_{j=1}^{k} 2^j + \sum_{j=2}^{k} 2^j + \ldots + \sum_{j=k}^{k} 2^j = \sum_{l=1}^{k} \sum_{j=l}^{k} 2^j \\
&= \sum_{l=1}^{k} \left(\sum_{j=1}^{k} 2^j - \sum_{j=1}^{l-1} 2^j \right) = \sum_{l=1}^{k} \left(2^{k+1} - 2 - (2^l - 2) \right) \\
&= k2^{k+1} - \sum_{l=1}^{k} 2^l = k2^{k+1} - (2^{k+1} - 2) = (k-1)2^{k+1} + 2 \qquad (531)
\end{aligned}
$$

...

Lempel-Ziv: Demonstração IV

Demonstração.

continuação...

Considere uma *string* binária de comprimento n. Observe que o número de frases distintas $c(n)$ desta *string* será maximizado quando todas as frases forem tão menores quanto possível.

- Vamos inicialmente considerar o caso em que $n = n_k$ (ou seja, vamos considerar uma *string* de comprimento n igual à soma dos comprimentos de todas as *strings* distintas de comprimento $\leq k$). Então c será maximizado quando considerarmos todas as *strings* distintas de comprimento $\leq k$. Iremos considerar $k \geq 2$, para que tenhamos efetivamente *strings*.

...

Lempel-Ziv: Demonstração V

Demonstração.

continuação...

O número de *strings* distintas de comprimento $\leq k$ é

$$c(n) = c(n_k) \leq \sum_{j=1}^{k} 2^j = 2^{k+1} - 2 < 2^{k+1} < \frac{n_k}{k-1} = \frac{n}{k-1}, \tag{532}$$

onde utilizamos $\sum_{j=1}^{k} 2^j = 2^{k+1}$, como visto anteriormente, e $\frac{n_k}{k-1} > 2^{k+1}$, pois

...

Lempel-Ziv: Demonstração VI

Demonstração.

continuação...

$$\begin{aligned} n_k &= (k-1)2^{k+1}+2 \\ n_k - 2 &= (k-1)2^{k+1} \\ n_k &> (k-1)2^{k+1} \\ \frac{n_k}{k-1} &> 2^{k+1} \end{aligned} \tag{533}$$

...

Lempel-Ziv: Demonstração VII

Demonstração.

continuação...

- Vamos agora considerar o caso em que $n \neq n_k$.

Note que:

$$n_k + (k+1)2^{k+1} = 2k2^{k+1} + 2 = k2^{k+2} + 2 = n_{k+1} \tag{534}$$

onde novamente utilizamos $n_k = (k-1)2^{k+1} + 2$.

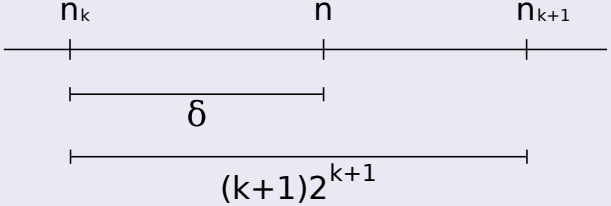
...

Lempel-Ziv: Demonstração VIII

Demonstração.

continuação...

- Supondo $n_k \leq n < n_{k+1}$, digamos que $n = n_k + \delta_k$ com $\delta_k < (k+1)2^{k+1} = n_{k+1} - n_k$ (que é o próximo passo de n_k).



...

Lempel-Ziv: Demonstração IX

Demonstração.

continuação...

- Considerando *strings* de comprimento $n_k \leq n < n_{k+1}$, ao realizar a análise (*parsing*) em frases distintas com menor comprimento possível, teremos $c(n)$ maximizado quando:
  1) não utilizarmos frases com comprimento maior que $k+1$;
  2) e utilizarmos todas as frases com comprimento $\leq k$, sendo que existem $c(n_k) < n/(k-1)$;
  3) as *strings* remanescentes, $(n - n_k) = \delta$, serão analisadas em frases únicas de comprimento $k+1$, então existe um total de $\delta/(k+1)$ dessas *strings*.

...

Lempel-Ziv: Demonstração X

Demonstração.

continuação...

Teremos então:

$$c(n) < \frac{n_k}{k-1} + \frac{\delta}{k+1} < \frac{n_k + \delta}{k-1} = \frac{n}{k-1}, \tag{535}$$

onde k é tal que $n_k \leq n < n_{k+1}$.

...

Lempel-Ziv: Demonstração XI

Demonstração.

continuação...

- para um dado n, vamos considerar $n_k \leq n < n_{k+1}$;
- vamos limitar k dado n (onde n é o comprimento da *string* e k é o comprimento da maior frase de análise);
- como $n \geq n_k = (k-1)2^{k+1} + 2 > 2^k$, teremos $k < \log n$

$$\begin{aligned} n &\geq n_k = (k-1)2^{k+1} + 2 \\ &= \underbrace{(k-1)}_{\geq 1} 2 2^k + 2 > 2^k. \end{aligned} \tag{536}$$

onde utilizamos que $k \geq 2$, como argumentado anteriormente.

...

Lempel-Ziv: Demonstração XII

Demonstração.

continuação...

- teremos também

$$\begin{aligned} n &< n_{k+1} = k2^{k+2} + 2 < (k+2)2^{k+2} \\ &< (\log n + 2)2^{k+2} \end{aligned} \tag{537}$$

onde utilizamos que $k < \log n$.

- teremos assim

$$k + 2 > \log\left(\frac{n}{\log n + 2}\right) \tag{538}$$

...

Lempel-Ziv: Demonstração XIII

Demonstração.

continuação...

- Para $n \geq 4$, iremos subtrair 3 de cada lado da Equação 538:

$$\begin{aligned} k-1 &> \log n - \log(\log n + 2) - 3 \\ &= \left(1 - \frac{\log(\log n + 2) + 3}{\log n}\right) \log n \\ &\geq \left(1 - \frac{\log(2 \log n) + 3}{\log n}\right) \log n \quad n \geq 4 \Rightarrow \log n \geq 2 \\ &= \left(1 - \frac{\log(\log n) + 4}{\log n}\right) \log n = (1 - \epsilon_n) \log n \end{aligned} \tag{539}$$

onde

$$\epsilon_n = \min\left\{1, \frac{\log \log n + 4}{\log n}\right\} \to 0. \tag{540}$$

...

Lempel-Ziv: Demonstração XIV

Demonstração.

continuação...

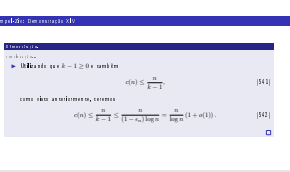
- Utilizando que $k-1 \geq 0$ e também

$$c(n) \leq \frac{n}{k-1}, \tag{541}$$

como visto anteriormente, teremos

$$c(n) \leq \frac{n}{k-1} \leq \frac{n}{(1-\epsilon_n)\log n} = \frac{n}{\log n}\left(1+o(1)\right). \tag{542}$$

□



$$\begin{aligned}
\sum_{l=0}^{k} 2^l &= 1 + 2 + 4 + \ldots + 2^k \\
(1-2)\left(\sum_{l=0}^{k} 2^l\right) &= (1 + 2 + 4 + \ldots + 2^k)(1-2) \\
(-1)\sum_{l=0}^{k} 2^l &= 1 - 2^{k+1} \\
\sum_{l=0}^{k} 2^l &= 2^{k+1} - 1 \qquad (543)
\end{aligned}$$

Distribuição de Máxima Entropia I

Teorema (Distribuição de Máxima Entropia)

Seja f uma função densidade probabilidade com suporte S (ou seja, $\int_S f(x)dx = 1$) satisfazendo as seguintes restrições de momento

$$\int_S f(x) r_i(x) dx = \alpha_i \quad , 1 \leq i \leq m. \tag{544}$$

A função de densidade da forma $f_\lambda(x) = e^{\lambda_0 + \sum_{i=1}^m \lambda_i r_i(x)}$, onde $\lambda = (\lambda_0, \ldots, \lambda_m)$ é escolhido de forma que $f_\lambda(x)$ satisfaça as restrições, é a única distribuição que satisfaz as restrições e maximiza a entropia diferencial $h(f)$.

- Demonstração: capítulo 12 (utiliza multiplicadores de Lagrange).
- A distribuição Gaussiana, para um dada média ($\mathrm{E}X = \mu$) e covariância ($\mathrm{E}XX^T = C$), é a única distribuição com entropia máxima.

Distribuição de Máxima Entropia II

Lema (limite da entropia)

Seja Z uma v.a. inteira positiva com média μ. Podemos limitar $H(Z)$ da seguinte forma:

$$H(Z) \leq (\mu+1)\log(\mu+1) - \mu\log\mu. \tag{545}$$

ideias para a demonstração.

- Uma v.a. Z com distribuição geométrica (definida nos inteiros $1, 2, \ldots$) com média $\mu = 1/p$ é definida como $\Pr(Z = k) = (1-p)^k p$.
- A entropia da distribuição geométrica com média μ é dada por $(\mu+1)\log(\mu+1) - \mu\log\mu$.
- A distribuição geométrica é a distribuição com máxima entropia dentre todas aquelas definidas nos inteiros positivos com média μ (mais uma vez utilizando os multiplicadores de Lagrande para demonstrar).

□

Ergodicidade I

- Processos ergódicos não podem ser separados em diferentes modos comportamentais persistentes.
- A média temporal é a mesma que a média amostral.
- Seja $x = \{x_i\}$ uma sequência de letras produzida pela fonte.
- $T^l x$ é a sequência deslocada no tempo de l posições. Se $T^l x = x'$, então $x'_i = x_{i+l}$, $\forall i$.
- Chamaremos de S um conjunto infinito de sequências de símbolos da fonte, isto é, $S = \{x : x \text{ é uma sequência de símbolos da fonte }\}$.
- $T^l S$ é o conjunto de todas as sequências deslocadas de l posições, isto é, se $x' = T^l x$, então $x' \in T^l S$ se $x \in S$.
- Um conjunto S é **invariante** se $T^l S = S$, $\forall l$.
  - exemplo: o conjunto de todas as sequências de uma fonte com alfabeto discreto é invariante.
  - exemplo: $S = \{\ldots 000 \ldots, \ldots 111 \ldots\}$ é invariante.

Ergodicidade II

- exemplo: para qualquer sequência x, o conjunto

$$\{\ldots, T^{-2}x, T^{-1}x, T^{0}x, T^{1}x, T^{2}x, \ldots\} \tag{546}$$

  é invariante (o conjunto de todas os possíveis deslocamentos de uma sequência x).

- Uma fonte estacionária discreta é **ergódica** se todo conjunto invariante de sequências possuir probabilidade 1 ou 0, ou seja,

$$\Pr\{S : T^{l}S = S, \forall l\} = 1 \text{ ou } 0, \quad \forall S. \tag{547}$$

Aproximação de Markov de ordem k I

- Seja $\{X_i\}_{i=-\infty}^{\infty}$ um processo estacionário ergódico com função massa de probabilidade $p(x_1, x_2, \ldots, x_n)$, então para um inteiro k fixo, definimos a aproximação de ordem k para p como

$$Q_k(\underbrace{x_{-(k-1)}, \ldots, x_0}_{\text{estado passado}}, x_1, \ldots, x_n) \triangleq p(x_{-(k-1):0}) \prod_{j=1}^{n} p(x_j | x_{j-k:j-1}) \tag{548}$$

  onde interpretamos $x_{-(k-1):0}$ como o estado '0', $x_{j-k:j-1}$ como o estado 'j' e $x_{i:j} \triangleq \{x_i, x_{i+1}, \ldots, x_j\}$ com $i \leq j$.

- Note que $p(x_n | x_{n-k:n-1})$ é estacionário e ergódico, visto que $p(x_{n-k:n})$ também é.

Aproximação de Markov de ordem k II

- Teremos que

$$\begin{aligned} -\frac{1}{n}\log Q_k(x_1,\ldots,x_n|x_{-(k-1):0}) &= -\frac{1}{n}\sum_{j=1}^{n}\log p(x_j|x_{j-k:j-1}) \\ \to -\mathrm{E}\log p(X_j|X_{j-k:j-1}) &= H(X_j|X_{j-k:j-1}) \end{aligned} \tag{549}$$

  já que o processo é estacionário ergódico.

- Iremos mostrar que

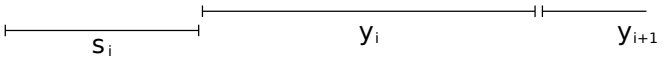
$$\limsup_{n\to\infty}\frac{c(n)\log c(n)}{n} \leq H(X_j|X_{j-k:j-1}) \to H(\mathcal{X}) \tag{550}$$

  onde $H(\mathcal{X})$ é a taxa de entropia do processo estocástico.

Notações I

- Uma *string* $x_{1:n}$, com $x_i \in \mathcal{X}$, será analisada (*parsed*) em c frases distintas, $x_{1:n} = y_1 y_2 \dots y_c$ onde y_i, para $i \in \{1, \dots, c\}$, é uma subsequência.
- Faremos v_i o índice em $x_{1:n}$ do início da i-ésima frase, ou seja, $y_i = x_{v_i:v_{i+1}-1}$ para todo $i = 1, 2, \dots, c$.
- Os k símbolos de $x_{1:n}$ que antecedem y_i serão chamados de s_i, ou seja, $s_i = x_{v_i-k:v_i-1}$. Para $i = 1$ teremos $s_1 = x_{-(k-1):0}$, então s_i é p 'estado' ou o prefixo da i-ésima frase.

x_1 x_2 ... x_{v_i-k} ... x_{v_i-1} x_{v_i} x_{v_i+1} ... $x_{v_{i+1}-1}$ $x_{v_{i+1}}$ ... x_{n-1} x_n



- Para $l \in \{1, 2, \dots, n\}$ e $s \in \mathcal{X}^k$ (uma *string* de comprimento k), vamos definir c_{ls} como o número de frases em $x_{1:n}$ com comprimento l e que possuem como estado anterior s, ou seja, c_{ls} é o número de frases de comprimento l precedidas por s, ou ainda

$$c_{ls} = |\{i : i \in \{1, \dots, c\}, |y_i| = l, s_i = s\}|. \tag{551}$$

Notações II

- Teremos assim $\sum_{l,s} c_{ls} = c = c(n)$, onde $c = c(n)$ é o número total de frases em uma análise em frases distintas de uma sequência de comprimento n.
- E ainda, $\sum_{l,s} l c_{ls} = n$, que é o comprimento total da *string*.
- Pergunta central: podemos relacionar (ou ao menos limitar) a probabilidade a algum aspecto determinístico de uma análise (*parsing*)? Veremos como isso é possível, e assim poderemos usar a análise baseada no *parsing* para impor um limite sobre a taxa de entropia do processo.

Lema (Desigualdade de Ziv)

Para qualquer análise (parsing) distinta (o que inclui a análise do LZ) de uma string $x_{1:n}$, temos:

$$\log Q_k(x_1, \ldots, x_n | s_1) \leq -\sum_{l,s} c_{ls} \log c_{ls} \tag{552}$$

Notações III

- Note que o limite é independente de Q e depende apenas de c_{ls}, que é o número de frases de comprimento l e antecedidas por s (estado anterior).
- Ideia principal: quanto maior for a diversidade em $x_{1:n}$, a maior probabilidade possível diminui, ou seja, y_i's distintas aumentam a diversidade.

Desigualdade de Ziv.

- Inicialmente temos

$$Q_k(x_{1:n}|s_1) = Q_k(y_{1:c}|s_1) = \prod_{i=1}^{c} p(y_i|s_i) \tag{553}$$

que segue ao assumir Markovidade de ordem k, ou seja, que y_i não depende de nada no passado, dado o passado imediato s_i.

...

Notações IV

Desigualdade de Ziv.

continuação...

- Temos então, tomando o $\log$ da equação anterior,

$$\begin{aligned}\log Q_k(x_1,\ldots.x_n|s_1) &= \sum_{i=1}^{c}\log p(y_i|s)\\ &= \sum_{l,s}\sum_{i:|y_i|=l,s_i=s}\log p(y_i|s_i)\\ &= \sum_{l,s}c_{ls}\sum_{i:|y_i|=l,s_i=s}\frac{1}{c_{ls}}\log p(y_i|s_i) \qquad (554)\end{aligned}$$

...

Notações V

Desigualdade de Ziv.

continuação...

- entretanto, temos uma mistura cujos coeficientes somam um, pois

$$\sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} = 1 \tag{555}$$

já que c_{ls} é o número de frases de comprimento l, antecedidas por s, e que as somas somam sobre todos os termos. Podemos assim ver esse valores $(1/c_{ls})$ como uma distribuição.

...

Notações VI

Desigualdade de Ziv.

continuação...

- Utilizando a desigualdade de Jensen para o termo em parênteses, teremos

$$\sum_{l,s} c_{ls} \left(\sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} \log p(y_i|s_i) \right) \leq \sum_{l,s} c_{ls} \log \left(\sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} p(y_i|s_i) \right) \tag{556}$$

- Mas todos os y_i's são distintos (não são contabilizados duplamente no somatório), e assim

$$\sum_{i:|y_i|=l, s_i=s} p(y_i|s_i) \leq \sum_{y'} p(y'|s) = 1 \tag{557}$$

...

Notações VII

Desigualdade de Ziv.

continuação...

- Teremos então o seguinte resultado, para qualquer distribuição,

$$\begin{aligned} \log Q_k(x_1, \ldots . x_n | s_1) &\leq \sum_{l,s} c_{ls} \log \left(\sum_{i:|y_i|=l, s_i=s} \frac{1}{c_{ls}} p(y_i|s_i) \right) \\ &\leq \sum_{l,s} c_{ls} \log \left(\frac{1}{c_{ls}} \right). \end{aligned} \tag{558}$$

Obtivemos um limite que depende exclusivamente da análise (*parsing*).

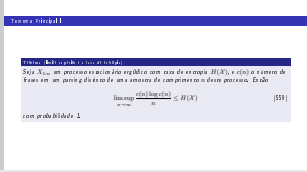
Teorema Principal I

Teorema (limite superior é a taxa de entropia)

Seja $X_{1:n}$ um processo estacionário ergódico com taxa de entropia $H(\mathcal{X})$, e $c(n)$ o número de frases em um parsing distinto de uma amostra de comprimento n deste processo. Então

$$\limsup_{n\to\infty} \frac{c(n)\log c(n)}{n} \leq H(\mathcal{X}) \tag{559}$$

com probabilidade 1.



$$\limsup_{n\to\infty} a_n \triangleq \inf_{n>0}\left(\sup_{k>n} a_k\right) = \inf S \tag{560}$$

onde $S = \{a : a = \sup_n B_n,\ \text{com}\ B_n = \{a_n, a_{n+1}, \ldots\}\}$.

- exemplo: $\lim_{x\to\infty} \sin(x)$ não existe, mas $\limsup_{x\to\infty} \sin(x) = 1$.

Temos também

$$\limsup_{n\to\infty} a_n \triangleq \sup_{n>0}\left(\inf_{k>n} a_k\right) \tag{561}$$

então, $\limsup$ requer a convergência do supremo no máximo local.

Teorema Principal - demonstração I

limite superior é a taxa de entropia.

Para simplificar escreveremos c para $c(n)$.

- Utilizando a desigualdade de Ziv temos

$$\begin{aligned} \log Q_k(x_1, \ldots, x_n | s_1) &\leq -\sum_{l,s} \frac{c_{ls} c}{c} \log \frac{c_{ls} c}{c} \\ &= -c \log c - c \sum_{l,s} \frac{c_{ls}}{c} \log \frac{c_{ls}}{c} \end{aligned} \tag{562}$$

- como $\sum_{l,s} c_{ls} = c$, vamos escrever $\pi_{ls} = c_{ls}/c$, que pode ser tratado como uma probabilidade, já que $\pi_{ls} \geq 0$ e $\sum_{l,s} \pi_{ls} = 1$.

...

Teorema Principal - demonstração II

limite superior é a taxa de entropia.

continuação...

- Como $\sum_{l,s} l c_{ls} = n$, temos

$$\sum_{l,s} l\pi_{ls} = n/c. \tag{563}$$

- Vamos definir duas v.a. U e V, tais que

$$p(U = l, V = s) = \pi_{ls} \tag{564}$$

de forma que

$$EU = \sum_{l} l\pi_{l} = \sum_{l} l \sum_{s} \pi_{ls} = n/c \tag{565}$$

...

Teorema Principal - demonstração III

limite superior é a taxa de entropia.

continuação...

- Isto nos leva imediatamente a

$$\begin{aligned} \log Q_k(x_{1:n}|s_1) &\leq \sum_{ls} c_{ls} \log 1/c_{ls} \\ &= -c \log c - c \sum_{l,s} \frac{c_{ls}}{c} \log \frac{c_{ls}}{c} \\ &= cH(U,V) - c \log c. \end{aligned} \tag{566}$$

...

Teorema Principal - demonstração IV

limite superior é a taxa de entropia.

continuação...

$$\underbrace{-\frac{1}{n}\log Q_k(x_{1:n}|s_1)}_{\substack{\to \text{ taxa de entropia quando} \\ k\to\infty \text{ e } n\to\infty}} \geq \underbrace{\frac{c}{n}\log c}_{\substack{c=c(n), \\ \text{queremos mostrar que} \\ \text{converge para entropia de } X}} \underbrace{-\frac{c}{n}H(U,V)}_{\substack{\text{idealmente } \to 0 \\ \text{quando } n\to\infty}} \tag{567}$$

- Sabemos que $H(U,V) \leq H(U) + H(V)$.
- E também $H(V) \leq \log|\{0,1\}|^k = k$. Podemos pensar em V como uma variável de estado (*string* binária de comprimento k).

...

Teorema Principal - demonstração V

limite superior é a taxa de entropia.

continuação...

- Utilizando o lemma anterior,

$$H(Z) \leq (\mu + 1)\log(\mu + 1) - \mu \log \mu, \tag{568}$$

teremos:

$$H(U) \leq (\mathrm{E}U + 1)\log(\mathrm{E}U + 1) - \mathrm{E}U \log \mathrm{E}U \tag{569}$$

...

Teorema Principal - demonstração VI

limite superior é a taxa de entropia.

Teorema Principal - demonstração VII

continuação...

$$
\begin{aligned}
H(U) &\leq (\mathrm{E}U+1)\log(\mathrm{E}U+1) - \mathrm{E}U \log \mathrm{E}U \\
&= \left(\frac{n}{c}+1\right)\log\left(\frac{n}{c}+1\right) - \frac{n}{c}\log\frac{n}{c} \\
&= \frac{n}{c}\log\left(\frac{n}{c}+1\right) + \log\left(\frac{n}{c}+1\right) - \frac{n}{c}\log\frac{n}{c} \\
&= \frac{n}{c}\log\left(\frac{c}{n}\left(\frac{n}{c}+1\right)\right) + \log\left(\frac{n}{c}+1\right) \\
&= \frac{n}{c}\log\left(\frac{c}{n}\times\frac{n}{c}+\frac{c}{n}\right) + \log\left(\frac{n}{c}+1\right) \\
&= \frac{n}{c}\log\left(\frac{c}{n}+1\right) + \log\left(\frac{n}{c}+1\right) + \log\left(\frac{c}{n}+1\right) - \log\left(\frac{c}{n}+1\right) \\
&= \left(\frac{n}{c}+1\right)\log\left(\frac{c}{n}+1\right) + \log\left(\frac{\frac{n}{c}+1}{\frac{c}{n}+1}\right) \qquad (570)
\end{aligned}
$$

...

limite superior é a taxa de entropia

Comprimentos I

Teorema

Seja X_i um processo estocástico estacionário de comprimento infinito. Seja $l(x_{1:n})$ os comprimentos das palavras LZ para n símbolos. Então

$$\limsup_{n\to\infty} \frac{1}{n} l(x_{1:n}) \leq H(\mathcal{X}) \tag{578}$$

Comprimentos II

Demonstração.

- Sabemos que $l(x_{1:n}) = c(n)(\log(c(n)) + 1)$, onde $c(n)$ é o número de frases da análise LZ (distintas).
- Pelo lemma anterior temos que

$$\limsup_{n\to\infty} \frac{c(n)}{n} = \limsup_{n\to\infty} \frac{1 + o(1)}{\log n} = 0 \tag{579}$$

- Portanto

$$\limsup_{n\to\infty} \frac{l(x_{1:n})}{n} = \limsup_{n\to\infty} \left(\underbrace{\frac{c(n)\log c(n)}{n}}_{\to H(\mathcal{X})} + \underbrace{\frac{c(n)}{n}}_{\to 0} \right) \leq H(\mathcal{X}). \tag{580}$$

□

Comprimentos III

- Em outras palavras, um algoritmo procedural (LZ), quando lida com um processo estocástico estacionário ergódico governado por uma dada distribuição, sem a necessidade de conhecer a distribuição e apenas seguindo os passos do algoritmo, irá convergir pra a taxa de entropia do processo estocástico, no limite.

Compressão I

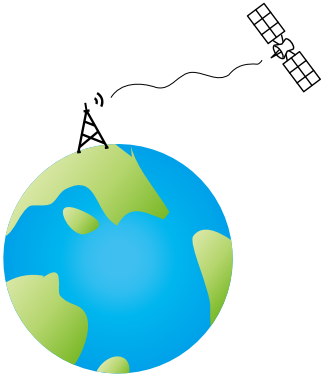
- Dada uma fonte com distribuição $p(x)$ e informação $H(X)$, vimos que a entropia é o limite da compressão. De certa forma, podemos dizer que a representação daquela informação possui uma determinada 'capacidade'. Com n bits podemos representar no máximo n bits de informação.
- Compressão é o processo de utilizar o máximo possível a capacidade de representação da informação produzida por uma fonte. Foi definido o conceito de eficiência: $H(X)/\mathrm{E}l$

Comunicação através de um canal I

- Claude Shannon

  Um problema fundamental em comunicação é reproduzir em um ponto exatamente ou aproximadamente uma mensagem selecionada em outro ponto.
  Frequentemente a mensagem possui algum significado... [o que é] irrelevante para o problema de engenharia.

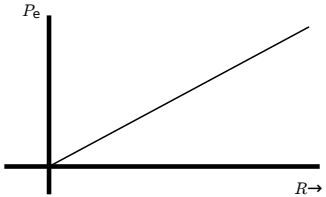
Comunicação através de um canal II



- Existe um limite para a taxa de comunicação em um canal?
- Se o canal possuir ruido, é possível realizar uma transmissão sem erro? A qual taxa?

Comunicação através de um canal III

- Existe um limite superior para a quantidade de informação que podemos enviar por um canal dependendo das condições de ruido presente?
- Aumentar a taxa de transmissão acarretará num aumento da probabilidade de erro?



- Se isto ocorrer, então a única forma de ter $P_e = 0$ seria não comunicar, $R = 0$.

Comunicação através de um canal IV

Exemplo (código de repetição)

- Vamos representar um sinal por uma sequência de números.
- Sabemos que um sinal pode ser reconstruído perfeitamente a partir de suas amostras (teorema da amostragem, Shannon).
- A sequência será transmitida por um canal AM ruidoso.
- Possivelmente algum número será mascarado pelo ruído.
- Podemos repetir cada número k vezes, escolhendo k grande suficiente, garantindo assim que seremos capaz de decodificar a sequência original com probabilidade de erro pequena.
- A probabilidade de error diminui com k e, ao mesmo tempo, a taxa de comunicação também diminui.

...

Comunicação através de um canal V

Exemplo (código de repetição)

continuação...

- suponha $k = 3$ e a probabilidade de erro p
- mensagem a ser transmitida pelo transmissor: $s = 10110$
- mensagem transmitida: $t = 111000111111000$.
- ruído: $n = 100011101010001$
- mensagem recebida: $r = t \oplus n$ (soma módulo 2)
- $r = 011011010101001$.
- decodificador: voto da maioria.

...

Comunicação através de um canal VI

Exemplo (código de repetição)

continuação...

- Probabilidade de inferência
- Regra do produto

$$\begin{aligned} \Pr(s,r) &= \Pr(s)\Pr(r|s) \\ &= \Pr(r)\Pr(s|r) \end{aligned} \tag{581}$$

- Regra da soma

$$\Pr(r) = \sum_s \Pr(s,r) = \Pr(s=0,r) + \Pr(s=1,r) \tag{582}$$

...

Comunicação através de um canal VII

Exemplo (código de repetição)

continuação...

- A probabilidade a posteriori de s é dada

$$\begin{aligned} \Pr(s|r) &= \frac{\Pr(r|s)\Pr(s)}{\Pr(r)} \\ &= \frac{\Pr(r|s)\Pr(s)}{\Pr(r|s=0)\Pr(s=0)+\Pr(r|s=1)\Pr(s=1)} \end{aligned} \tag{583}$$

- $\Pr(r|s)$: verossimilhança de s
- $P(s)$: probabilidade a priori de s

...

Comunicação através de um canal VIII

Exemplo (código de repetição)

continuação...

- suponha $r = 011$

$$\Pr(r|s=0) = (1-p) \times p \times p = (1-p)p^2 \tag{584}$$

$$\Pr(r|s=1) = p \times (1-p) \times (1-p) = p(1-p)^2 \tag{585}$$

- temos então

$$p^{\text{numero de trocas}}(1-p)^{\text{numero de concordancias entre t(s) e r}} \tag{586}$$

...

Comunicação através de um canal IX

Exemplo (código de repetição)

continuação...

- supondo $\Pr(s=0) = \Pr(s=1) = 1/2$, teremos

$$\Pr(s=1|r=011) = \frac{(1-p)^2 p \frac{1}{2}}{(1-p)p^2\frac{1}{2} + p(1-p)^2\frac{1}{2}} = (1-p) \tag{587}$$

- para $p = 0.1$, por exemplo, teremos probabilidade de 90%.
- $\Pr(s=1|r) > \Pr(s=0|r)$ então $\hat{s} = 1$ é o melhor chute.

Ideia Central I

- Se escolhermos as mensagens apropriadamente, podemos ter uma situação em que teremos uma alta probabilidade de ter identificação univoca destas mensagens no receptor.
- A ideia é escolher mensagens que tendem a não criar ambiguidade no receptor.

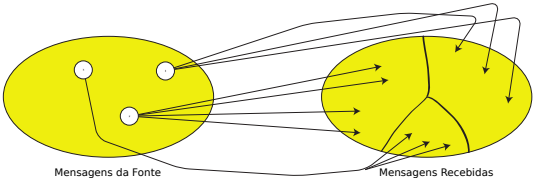


Figura 15: Evitar ambiguidades (Bilmes, 2013).

- Não vamos enviar o sinal, mas sim uma representação (ou uma descrição, ou uma codificação) do sinal que seja adequada ao canal, de forma que o receptor seja capaz de recuperar (decodificar) o sinal. Adicionaremos redundância de acordo com o canal. A P_e não precisa tender a zero.

Canal Discreto I

Definição (Canal Discreto)

Uma canal discreto é aquele em que existe um alfabeto de entrada $\mathcal{X}$, um alfabeto de saída $\mathcal{Y}$ e uma distribuição $p(y \mid x)$ que fornece a probabilidade de se observar uma saída y quando a entrada é x.

Definição (Canal sem memória)

Uma canal discreto é sem memória se y_t, a saída no instante t, é independente de todas entradas passadas, dado x_t. Isto é, $y_t \perp\!\!\!\perp x_{1:t-1} \mid x_t$.

Canal Discreto II

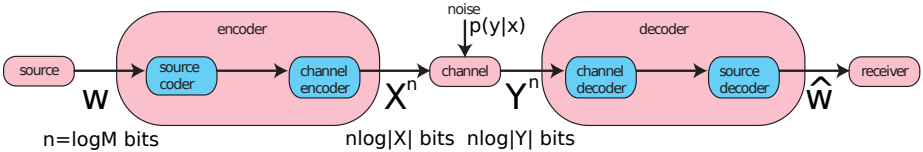


Figura 16: Modelo de Comunicação (Bilmes, 2013).

- $p(x, y) = p(x)p(y|x)$
- $p(y|x)$ modela o canal sendo fixo na maioria dos casos (não temos controle sobre ele).
- $p(x)$ é a distribuição da fonte, que pode ser modificada (otimizada).

Canal Discreto III

- $p(x)$ e $p(y|x)$ são suficientes para calcularmos a informação mutua entre X e Y.

$$\begin{aligned} I(X;Y) = I_{p(x)}(X;Y) &= \sum_{x,y} p(x)p(y|x)\log\frac{p(y|x)}{p(y)} \\ &= \sum_{x,y} p(x)p(y|x)\log\frac{p(y|x)}{\sum_{x'} p(y|x')p(x')} \end{aligned} \tag{588}$$

- Para um canal fixo ($p(y|x)$ fixo), a informação mútua é uma função da distribuição $p(x)$, $I(X;Y) = I_{p(x)}(X;Y)$. Esta função é concava em $p(x)$ para $p(y|x)$ dado (como visto anteriormente).
- Otimizar sobre $p(x)$, para $p(y|x)$ fixo, encontrando assim a informação mútua máxima.

Definição (fluxo de informação)

A taxa de fluxo de informação através de um canal é dada por $I(X;Y)$, em unidades de bits por utilização do canal.

Canal Discreto IV

Definição (capacidade)

A capacidade de informação de uma canal é o máximo de fluxo de informação que podemos ter neste canal.

$$C \triangleq \max_{p(x)\in\Delta} I(X;Y) \tag{589}$$

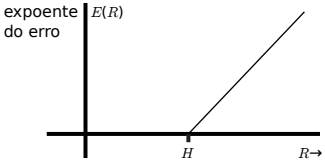
onde Δ é o conjunto de todas as distribuições de probabilidade sobre o alfabeto da fonte $\mathcal{X}$. Então C é o máximo de bits que podem ser enviados através do canal por utilização do canal.

Definição (taxa)

A taxa R de um código é medida em número de bits por utilização do canal.

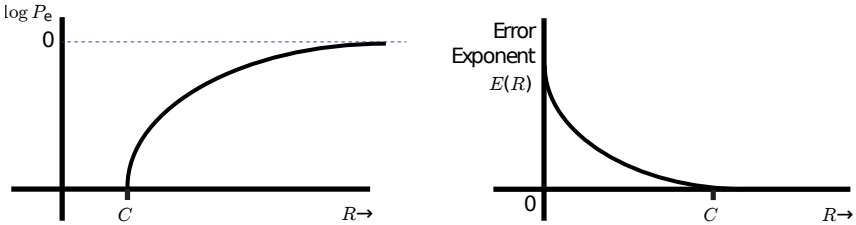
Canal Discreto I

- Para compressão, temos que $P_e \propto e^{-nE(R)}$. Se o expoente do erro é positivo, então erro $\rightarrow 0$ exponencialmente rápido à medida que o comprimento do bloco $\rightarrow \infty$.
- Para reduzir erro devemos ter $R > H$. Não é possível comprimir abaixo da entropia sem incorrer em erro.



- Shannon mostrou que algo semelhante ocorre em comunicação. Acima de uma quantidade fundamental C, à medida que a taxa R aumenta, acima de C, a probabilidade de erro cresce. Note que $P_e \propto e^{-nE(R)}$.

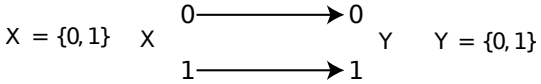
Canal Discreto II



- Iremos mostrar que a única maneira de ter pouco erro é se $R < C$.
- Note que podemos ter comunicação sem erro se $R < C$, sendo que $C > 0$.

Exemplo: Canal Discreto Sem Memória I

- Canal binário sem ruído de $\mathcal{X} = \{0, 1\}$ em $\mathcal{Y} = \{0, 1\}$. O diagrama mostra $p(y|x)$.



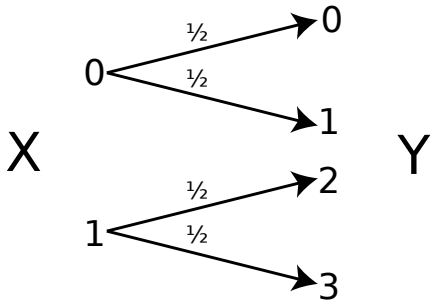
- Para este canal temos
  - $p(y = 0|x = 0) = 1 = 1 - p(y = 1|x = 0)$
  - $p(y = 1|x = 1) = 1 = 1 - p(y = 0|x = 1)$

  ou seja, a saída é uma cópia da entrada.
- Capacidade de 1 bit. Cada vez que enviamos um bit, ele será recebido do outro lado sem erro.
- $I(X;Y) = H(X) - H(X|Y) = H(X)$, neste caso, teremos

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} H(X) = 1 \tag{590}$$

- Claramente, a capacidade será alcançada se $p(0) = p(1) = 1/2$.
- Se $p(0) = 1 = 1 - p(1)$ teremos $I(X;Y) = 0$, e assim não teremos fluxo de informação.

Exemplo: Canal com ruído e saídas sem sobreposição I

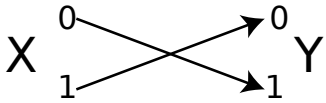


- Para este canal, temos
  - $p(Y = 0|X = 0) = p(Y = 1|X = 0) = 1/2$
  - $p(Y = 2|X = 1) = p(Y = 3|X = 1) = 1/2$
- Se recebermos 0 ou 1, sabemos que 0 foi enviado. Se recebermos 2 ou 3, sabemos que 1 foi enviado.

Exemplo: Canal com ruído e saídas sem sobreposição II

- $C = 1$
- $I(X;Y) = H(X) - \underbrace{H(X|Y)}_{=0} = H(X)$.
- $p(0) = p(1) = 1/2$ atingirá a entropia máxima.

Exemplo: Canal de permutação I

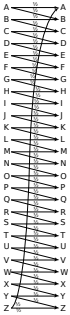


- Para este canal temos
  - $p(Y = 1|X = 0) = p(Y = 0|X = 1) = 1$
- Canal de permutação.
- $C = 1$
- De forma geral, para um alfabeto de tamanho $k = |\mathcal{X}| = |\mathcal{Y}|$, seja σ uma permutação de forma que $Y = \sigma(X)$, então $C = \log k$.

Otimização para calcular C I

- Para maximizar uma função $f(x)$, é suficiente mostrar que $f(x) \leq \alpha$ para todo x e então encontrar x^* tal que $f(x^*) = \alpha$.
- x^* atinge o limite superior α para $f(\cdot)$.
- $C = \max_{p(x)} I(X;Y)$ para $p(y|x)$ fixo.
- A solução $p^*(x)$ não é necessariamente única e não será necessariamente aquela escolhida para fazer a codificação de canal.
- C é resultado de uma otimização.
- C é o ponto crítico para sermos capazes de codificar para o canal com probabilidade de erro tendendo a zero.

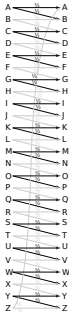
Máquina de Escrever com ruído I



- 26 símbolos
- cada símbolo é mapeado nele mesmo ou no seu vizinho
- i.e. $p(A \to A) = p(A \to B) = 1/2$, etc.

Máquina de Escrever com ruído II

- É possível comunicar sem erro selecionando um subconjunto que não gerará ambiguidade. Escolhendo A, C, E, ... ou Z, B, D, ...
- $A \to \{A, B\}$, $C \to \{C, D\}$, $E \to \{E, F\}$, etc.



- $C = \log 13$

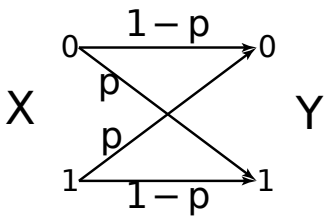
Máquina de Escrever com ruído III

- Matematicamente

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max_{p(x)} \left(H(Y) - H(Y|X)\right) \\ &= \max_{p(x)} H(Y) - 1 \quad \text{para } X = x, \exists \text{ duas opções} \\ &= \log 26 - 1 = \log 13 \end{aligned} \tag{591}$$

- $\max_{p(x)} H(Y) = \log 26$ pode ser alcançado quando $p(x)$ é uniforme, neste caso para qualquer x, teremos igual probabilidade de receber um dos dois Ys possíveis.
- Outra alternativa é escolher $p(x)$ de forma que a probabilidade seja nula em entradas alternadas (B, D, F, etc.). Neste caso, teremos também $H(Y) = \log 26$.
- A capacidade é a mesma em ambos casos, mas apenas um deles será utilizado para realizar uma codificação sem erro.

Canal Binário Simétrico I



- Um bit enviado poderá ser alterado (*flip*) com probabilidade p.
- O canal é caracterizado por
  - $p(Y = 1|X = 0) = p = 1 - p(Y = 0|X = 0)$
  - $p(Y = 0|X = 1) = p = 1 - p(Y = 1|X = 1)$
- É possível realizar comunicação sem erro com este canal? Sim, se a taxa de transmissão não for muito alta ($R > C$).

Canal Binário Simétrico II

- Calcular a capacidade

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(Y) - \sum_x p(x) H(Y|X=x) \\ &= H(Y) - \sum_x p(x) H(p) = H(Y) - H(p) \\ &\leq 1 - H(p) \end{aligned} \tag{592}$$

- Para alcançar o limite superior, precisamos $H(Y) = 1$, ou seja, $Y \sim \mathcal{U}$, $\Pr(Y=1) = \Pr(Y=0) = 1/2$.

Canal Binário Simétrico III

- Temos que

$$\begin{aligned}\Pr(Y=1) &= \Pr(Y=1|X=1)\Pr(X=1) \\ &\quad + \Pr(Y=1|X=0)\Pr(X=0) \\ &= (1-p)\Pr(X=1) + p(1-\Pr(X=1)) \\ &= p + (1-2p)\Pr(X=1). \end{aligned} \tag{593}$$

- Queremos $\Pr(Y=1) = \Pr(Y=0)$, ou seja,

$$\begin{aligned}\Pr(Y=0) &= \Pr(Y=1) \\ 1-p-(1-2p)\Pr(X=1) &= p + (1-2p)\Pr(X=1) \\ 1-2p &= 2(1-2p)\Pr(X=1) \\ 1 &= 2\Pr(X=1) \\ \frac{1}{2} &= \Pr(X=1). \end{aligned} \tag{594}$$

Canal Binário Simétrico IV

- Se $\Pr(X = 1) = 1/2$ teremos $\Pr(Y = 1) = p + 1/2 - p = 1/2$.
- Então $H(Y) = 1$ se $H(X) = 1$ (i.e. se $\Pr(X = 1) = 1/2$).
- $C = 1 - H(p)$ quando X é uniforme.
- Se $p = 1/2$ então $C = 0$, o canal irá alterar os bits aleatoriamente com igual probabilidade, e desta forma não será possível enviar informação alguma.
- Se $p \neq 1/2$, será possível comunicar, embora potencialmente devagar. Por exemplo, se $p = 0.499$, então $C = 2.8854 \times 10^{-6}$ bits por utilização do canal. Para enviar um único bit precisaremos utilizar o canal muitas vezes.
- Se $p = 0$ ou $p = 1$, teremos $C = 1$ e teremos assim a maior taxa possível (i.e. um bit por utilização do canal).

Decodificação I

- Podemos 'decodificar' a mensagem enviada pela fonte a partir da mensagem recebida, da distribuição da fonte e do modelo do canal $p(y|x)$ utilizando a regra de Bayes.

$$\Pr(x|y) = \frac{\overbrace{\Pr(y|x)}^{\text{canal}}\overbrace{\Pr(x)}^{\text{fonte}}}{\Pr(y)} = \frac{\Pr(y|x)\Pr(x)}{\sum_{x'}\Pr(y|x')\Pr(x')} \tag{595}$$

- Ao receber um determinado y, podemos calcular $p(x|y)$ e tomar uma decisão com base nisto. Ou seja, $\hat{x} = \operatorname{argmax}_x p(x|y)$ (decodificação de máxima verosimilhança).
- Ocorrerá um erro sempre que $\hat{x} \neq x$, então $\Pr(\text{erro}) = \Pr(x \neq \hat{x})$.
- Esta decodificação é a decodificação ótima, no sentido de minimizar o erro, quando y é recebido para um determinado x enviado,

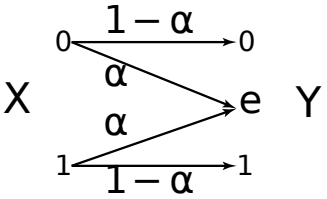
$$\text{erro}(\overline{x}) = 1 - p(\overline{x}|y(x)) \tag{596}$$

- Este erro será mínimo se escolhermos aquele que $\operatorname{argmax}_x p(x|y)$.

Decodificação com erro mínimo I

- Computar $\Pr(x|y)$ é uma tarefa de inferência probabilística.
- Este problema usualmente é difícil (NP-difícil). Isto implica que realizar a decodificação de erro mínimo pode ter um custo exponencial (a menos que P = NP).
- Para muitos métodos de codificação, o cálculo é realizado de maneira aproximada, uma vez que não se conhece nenhum método rápido de calcular o erro mínimo ou realizar a decodificação de máxima verossimilhança.
- Métodos de inferência aproximada, por exemplo, *loopy belief propagation*, *message passing*, etc., estes algoritmos tendem a funcionar muito bem na prática (atingem próximo de C).

Canal Binário com Apagamento I



- Neste exemplo temos um símbolo que representa que o símbolo transmitido foi perdido (apagado): e.
- A probabilidade disso acontecer é α.
- A capacidade é calculada da seguinte forma

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max_{p(x)} \left(H(Y) - H(Y|X)\right) \\ &= \max_{p(x)} H(Y) - H(\alpha) \end{aligned} \tag{597}$$

Canal Binário com Apagamento II

- Temos que $H(Y) \leq \log 3$.
- Seja a v.a. binária $E = \{Y = e\}$, então

$$H(Y) = H(Y, E) = H(E) + H(Y|E) \tag{598}$$

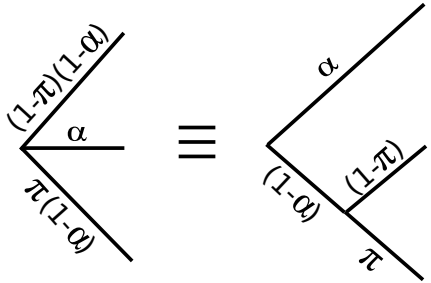
onde utilizamos a definição de E e a regra da cadeia.

- Seja $\pi = \Pr(X = 1)$, então

$$\begin{aligned} H(Y) &= H\left(\overbrace{(1-\pi)(1-\alpha)}^{\text{se } Y=0}, \overbrace{\alpha}^{\text{se } Y=e}, \overbrace{\pi(1-\alpha)}^{\text{se } Y=1}\right) \\ &= H(E) + H(Y|E) \\ &= H(\alpha) + (1-\alpha)H(\pi) \end{aligned} \tag{599}$$

Canal Binário com Apagamento III

- Na ultima igualdade utilizamos $H(E) = H(\alpha)$, e $H(Y|E) = \alpha H(Y|Y = e) + (1 - \alpha)H(Y|Y \neq e) = \alpha \cdot 0 + (1 - \alpha)H(\pi)$



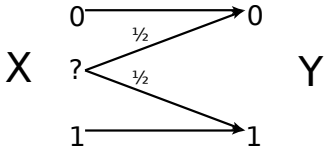
Canal Binário com Apagamento IV

- Teremos então

$$\begin{aligned} C &= \max_{p(x)} H(Y) - H(\alpha) \\ &= \max_{p(x)} \left((1-\alpha)H(\pi) + H(\alpha)\right) - H(\alpha) \\ &= \max_{p(x)} (1-\alpha)H(\pi) = 1-\alpha \end{aligned} \tag{600}$$

- Onde utilizamos que o máximo será obtido quando $\pi = 1/2 = \Pr(X=1) = \Pr(X=0)$.
- Neste canal perderemos α% dos bits transmitidos. A capacidade será máxima quando $\alpha = 0$, neste caso não haverá apagamento.

Canal de Confusão Ternário I



- $\Pr(Y = 0|X =?) = \Pr(Y = 1|X =?) = 1/2$
- Quando a entrada é ? teremos uma saída aleatória. As demais entradas são confiáveis.
- $C = 1$ bit.

Canal de Confusão Ternário II

- podemos obter da seguinte forma

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max_{p(x)} \left(H(Y) - H(Y|X)\right) \\ &= \max_{p(x)} \left(H(Y) - \Pr(X=?)\right) \\ &= 1 - 0 = 1 \end{aligned} \tag{601}$$

onde utilizamos que

$$\begin{aligned} H(Y|X) &= \Pr(X=0) \underbrace{H(Y|X=0)}_{=0} + \Pr(X=?) \underbrace{H(Y|X=?)}_{=1} + \\ &\quad \Pr(X=1) \underbrace{H(Y|X=1)}_{=0} \\ &= \Pr(X=?). \end{aligned} \tag{602}$$

Canal Simétrico I

Definição

Um canal é simétrico se as linhas da matriz de transmissão $p(y|x)$ são permutação uma das outras, e as colunas desta matriz também são permutação uma das outras. Um canal é fracamente simétrico se cada linha da matriz for uma permutação das outras linhas e todas as colunas tiverem a mesma soma $\sum_x p(y|x)$.

Teorema

Para um canal simétrico fraco, teremos

$$C = \log |\mathcal{Y}| - H(r) \tag{603}$$

onde r é a linha da matriz de transmissão.

O teorema acima segue do seguinte fato

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - H(r) \leq \log |\mathcal{Y}| - H(r) \tag{604}$$

Propriedades da Capacidade de Canal I

- $C \geq 0$, uma vez que $I(X;Y) \geq 0$,
- $C \leq \log |\mathcal{X}|$, uma vez que $C = \max_{p(x)} I(X;Y) \leq \max H(X) = \log |\mathcal{X}|$;
- de forma similar, teremos que $C \leq \log |\mathcal{Y}|$. Logo, o tamanho do alfabeto limita a capacidade de canal.
- $C \leq \log \left[\min \left(|\mathcal{X}|, |\mathcal{Y}|\right)\right]$
- $I(X;Y) = I_{p(x)}(X;Y)$ é uma função contínua de $p(x)$, logo, existe derivada e podemos otimizá-la.
- $I(X;Y)$ é uma função concava de $p(x)$ para $p(y|x)$ fixo.

$$I_{\lambda p_1 + (1-\lambda) p_2}(X;Y) \geq \lambda I_{p_1}(X;Y) + (1-\lambda) I_{p_2}(X;Y) \tag{605}$$

- Isto faz com que o cálculo da capacidade seja mais fácil, i.e., um máximo local é um máximo global, e calcular a capacidade para um modelo genérico de canal é um problema de otimização convexo.
- Temos também que $I(X;Y)$ é uma função concava de $p(y|x)$ para $p(x)$ fixo.

Segundo Teorema de Shannon I

Teorema (Segundo Teorema de Shannon)

C é o número máximo de bits (em média, por utilização do canal) que podemos transmitir através de um canal de forma confiável.

- Este teorema é um dos teoremas mais importantes do século XX.
- 'confiável' significa: com probabilidade de erro exponencialmente decrescente à medida que o comprimento do bloco cresce. Podemos fazer esta probabilidade essencialmente igual a zero.
- Por outro lado, se tentarmos enviar mais do que C bits através do canal, a probabilidade de erro rapidamente vai a 1.
- Problema de empacotamento de compartimentos. Temos uma região de possíveis palavras e iremos empacotar tantas quantas possível, de forma a criar compartimentos sem sobreposição.

Segundo Teorema de Shannon II

- Empacotamento em compartimentos não é um particionamento, uma vez que podemos ter espaço inutilizados.

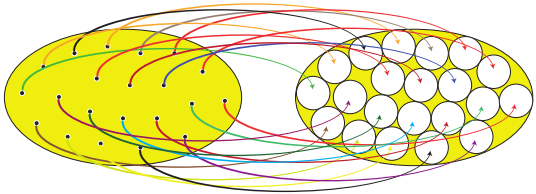


Figura 17: Empacotamento (Bilmes, 2013).

- Ideia intuitiva: utilizar tipicidade.
- Existem $\approx 2^{nH(X)}$ sequências típicas, cada uma com probabilidade $2^{-nH(X)}$ e $p(A_\epsilon^{(n)}) \approx 1$, então o conjunto típico possui 'toda' a probabilidade.

Segundo Teorema de Shannon III

- O mesmo ocorre para a entropia condicional, ou seja, para uma entrada típica X, existem $\approx 2^{nH(Y|X)}$ sequências de saída.
- Existem $2^{nH(Y)}$ sequências de saída típicas, e sabemos que $2^{nH(Y)} \geq 2^{nH(Y|X)}$.
- Objetivo é encontrar um subconjunto de entradas que não gerem confusão, ou seja, que produzam sequências na saída de forma disjunta (como ilustrado).
- Existem $\approx 2^{nH(Y)}$ saídas típicas (i.e., sequências em Y marginalmente típicas).
- Existem $\approx 2^{nH(Y|X)}$ saídas possíveis dadas as entradas possíveis (sequências típicas em Y condicionadas em X), i.e., o número médio de saídas para uma possível entrada, que serão confundidas entre si, ou seja, na média, para um dado $X = x$, será o número aproximado de saídas correspondentes que podem existir.
- O número de entradas inconfundíveis será

$$\leq \frac{2^{nH(Y)}}{2^{nH(Y|X)}} = 2^{n(H(Y)-H(Y|X))} = 2^{nI(X;Y)} \tag{606}$$

Segundo Teorema de Shannon IV

- Note que em uma situação não ideal, poderia haver sobreposição das sequências típicas em Y dado X, mas a melhor situação (maximizando as entradas não confundíveis) é quando não há sobreposição.

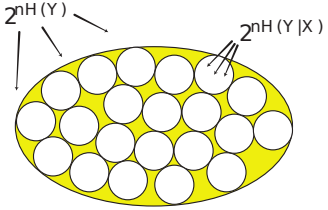


Figura 18: Empacotamento sem sobreposição (Bilmes, 2013).

- Para maximizar o número de entradas inconfundíveis, para um canal fixo ($p(y|x)$ fixo), devemos encontrar $p(x)$ que fornece $I(X;Y) = C$, que é o logaritmo do número máximo de entradas possíveis para utilização. C é a capacidade do canal.

Modelos de Comunicação I

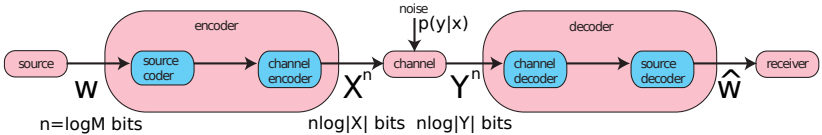


Figura 19: Modelo de comunicação (Bilmes, 2013).

Mensagem : $W \in \{1, \ldots, M\}$, sendo necessário $\log M$ bits por mensagem.

Sinal $X^n(W)$ (uma palavra aleatória) será enviado através do canal.

Sinal recebido através do canal, $Y^n \sim p(y^n|x^n)$.

Decodificação através de um chute $\hat{W} = g(Y^n)$.

Canal discreto sem memória : $(\mathcal{X}, p(y|x), \mathcal{Y})$.

n-ésima extensão do canal : $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$.

Código (M, n) I

Definição (Código (M, n))

Um código (M, n) para um canal $(\mathcal{X}, p(y|x), \mathcal{Y})$ é

1) um conjunto de índice $\{1, 2, \ldots, M\}$;
2) uma função de codificação $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$ levando a palavras $X^n(1), X^n(2), \ldots, X^n(M)$. Cada mensagem da fonte possui uma palavra (*codeword*) e cada palavra é um código de n símbolos.
3) função de decodificação, i.e., $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$ que realiza um 'chute' (adivinhação) sobre qual era a mensagem original dada a saída do canal.

- M é o número de possíveis mensagens a serem enviadas, e n é o número de utilizações do canal feitas pela palavra do código adotado.
- A taxa de comunicação é dada por $R = \log M/n$.

Erro I

Definição (Probabilidade de erro λ_i para a mensagem $i \in \{1, \ldots, M\}$)

$$\lambda_i \triangleq \Pr(g(Y^n) \neq i \mid X^n = x^n(i)) = \sum_{y^n \in \mathcal{Y}^n} p(y^n \mid x^n(i)) \mathbf{1}_{\{g(y^n) \neq i\}} \tag{607}$$

Definição (Probabilidade Máxima de Erro $\lambda^{(n)}$ para um código (M, n))

$$\lambda^{(n)} \triangleq \max_{i \in \{1,2,\ldots,M\}} \lambda_i \tag{608}$$

Erro II

Definição (Probabilidade de Erro Média $P_e^{(n)}$)

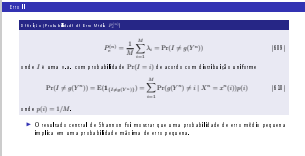
$$P_e^{(n)} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i = \Pr(I \neq g(Y^n)) \tag{609}$$

onde I é uma v.a. com probabilidade $\Pr(I = i)$ de acordo com distribuição uniforme

$$\Pr(I \neq g(Y^n)) = \mathrm{E}(\mathbf{1}_{\{I \neq g(Y^n)\}}) = \sum_{i=1}^{M}\Pr(g(Y^n) \neq i \mid X^n = x^n(i))p(i) \tag{610}$$

onde $p(i) = 1/M$.

- O resultado central de Shannon foi mostrar que uma probabilidade de erro médio pequena implica em uma probabilidade máxima de erro pequena.


Erro II

Teoria da Informação Capacidade de Canal Definições Erro

$$P_e^{(n)} = \frac{1}{M}\sum_{i=1}^{M}\lambda_i = \Pr(I \neq g(Y^n)) \quad (13)$$

onde I é uma v.a. com probabilidade $\Pr(I = i)$ de acordo com distribuição uniforme

$$\Pr(I \neq g(Y^n)) = \mathrm{E}(\mathbb{1}_{(I \neq g(Y^n))}) = \sum_{i=1}^{M} \Pr(g(Y^n) \neq i \mid X^n = x^n(i))p(i) \quad (14)$$

onde $p(i) = 1/M$.

- O resultado central de Shannon foi mostrar que uma probabilidade de erro média pequena implica em uma probabilidade máxima de erro pequena.

$P_e^{(n)}$, como definido acima, é apenas um construto matemático da probabilidade de erro condicional λ_i, sendo ele mesmo uma probabilidade de erro apenas se as mensagens forem escolhidas uniformemente sob o conjunto {1,2, . . . , M }. Escolhemos uma distribuição uniforme para achar um limite no erro, permitindo analisar o comportamento de $P_e^{(n)}$ e da probabilidade máxima de erro $\lambda^{(n)}$, caracterizando assim o comportamento do canal independente de como é utilizado (isto, para qualquer distribuição sobre W).

Taxa I

Definição (Taxa R de um código (M, n))

$$R = \frac{\log M}{n} \tag{611}$$

número total de bits em uma mensagem da fonte / número total de utilizações do canal necessárias para enviar a mensagem

Definição (Alcançabilidade de um dado canal)

Uma taxa R é alcançável para um dado canal se $\exists$ uma sequência de códigos $(\lceil 2^{nR} \rceil, n)$ tais que a probabilidade de erro máxima $\lambda^{(n)} \to 0$ quando $n \to \infty$.

Taxa I

Definição (Capacidade do canal discreto sem memória)

A capacidade de uma canal discreto sem memória é a maior taxa alcançável.

- A capacidade de um canal discreto sem memória é a taxa além da qual o erro não irá mais a zero com o crescimento de n.
- Note que esta definição (capacidade do canal discreto sem memória) é diferente da definição utilizada anteriormente ($C = \max_{p(x)} I(X;Y)$, a capacidade de informação).

Tipicidade Conjunta I

Definição (Tipicidade Conjunta de um conjunto de sequencias)

Um conjunto de sequências $\{(x_{1:n}, y_{1:n})\}$ com relação a $p(x, y)$ é tipicamente conjunto ($\in A_\epsilon^{(n)}$) segundo a definição a seguir:

$$
\begin{aligned}
A_\epsilon^{(n)} \quad = \quad & \{(x_{1:n}, y_{1:n}) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
& \text{a)} \quad \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \quad \text{típico em x} \\
& \text{b)} \quad \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \quad \text{típico em y} \\
& \text{c)} \quad \left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \epsilon, \quad \text{típico em (x,y)} \\
& \}
\end{aligned} \tag{612}
$$

com $p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$.

Tipicidade Conjunta II

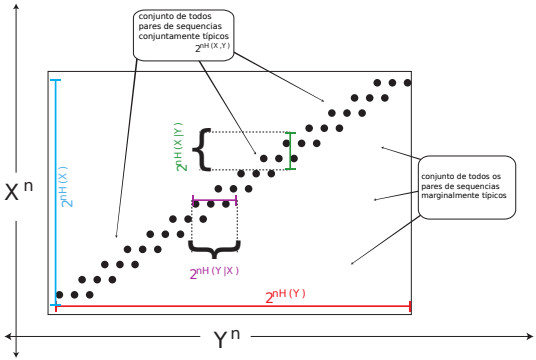


Figura 20: Sequências típicas (Bilmes, 2013).

Tipicidade Conjunta III

- razão entre o número de sequências conjuntamente típicas e o número de sequências típicas escolhidas de forma independente

$$\begin{aligned}\frac{\text{num. seq. conj. tip.}}{\text{num. seq. tip. escolhidas ind.}} &= \frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} \\ &= 2^{n(H(X,Y)-H(X)-H(Y))} \\ &= 2^{-nI(X;Y)} \end{aligned} \tag{613}$$

  a probabilidade de algo que seja marginalmente típico e conjuntamente típico decresce exponencialmente.
- Se escolhermos independentemente ao acaso duas sequências marginalmente típicas, para X e Y, então a probabilidade de que teremos um sequência típica conjuntamente em (X, Y) decrescerá exponencialmente com n, enquanto $I(X;Y) > 0$.
- Quando reduzir esta chance o máximo possível, teremos 2^{nC}.

Tipicidade Conjunta IV

Teorema (Prop. Equipartição Assintótica Conjunta)

Seja $(X^n, Y^n) \sim p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$. *Então*

1) $\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1$ *quando* $n \to \infty$.
2) $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$ *e* $(1-\epsilon)2^{n(H(X,Y)-\epsilon)} \leq |A_\epsilon^{(n)}|$.
3) *Se* $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ *são tirados de forma independente, então*

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)} \tag{614}$$

e para n *suficientemente grande, teremos*

$$\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \tag{615}$$

Tipicidade Conjunta V

Os limites na probabilidade de sequências retiradas de forma independente serem conjuntamente típicas cai exponencialmente rápido com n se $I(X;Y) > 0$.

Tipicidade Conjunta VI

$$\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1.$$

Tipicidade Conjunta VII

Pela lei fraca dos grandes números temos o seguinte

$$-\frac{1}{n}\log \Pr(X^n) \to -\mathrm{E}(\log p(X)) = H(X) \tag{616}$$

então $\forall \epsilon > 0$, $\exists m_1$ tal que para $n > m_1$

$$\Pr\left(\underbrace{|-\frac{1}{n}\log \Pr(X^n) - H(X)| > \epsilon}_{=S_1}\right) < \epsilon/3 \tag{617}$$

(apenas uma forma de reescrever o significado de convergência)

- Definimos S_1, o evento não típico.

...

Tipicidade Conjunta VIII

$\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1.$

continuação...

- O mesmo faremos com relação a Y e (X, Y).
- $\exists m_2$ tal que $\forall n > m_2$ teremos

$$\Pr\left(\underbrace{|-\frac{1}{n}\log\Pr(Y^n) - H(Y)| > \epsilon}_{=S_2}\right) < \epsilon/3 \tag{618}$$

...

Tipicidade Conjunta IX

$\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1.$

continuação...

- $\exists m_3$ tal que $\forall n > m_3$ teremos

$$\Pr\left(\underbrace{|-\frac{1}{n}\log\Pr(X^n, Y^n) - H(X,Y)| > \epsilon}_{=S_3}\right) < \epsilon/3 \tag{619}$$

- S_1, S_2 e S_3 são eventos não típicos.

...

Tipicidade Conjunta X

$\Pr\left((X^n, Y^n) \in A_\epsilon^{(n)}\right) \to 1.$

continuação...

- Para $n > \max(m_1, m_2, m_3)$, temos que $p(S_1 \cup S_2 \cup S_3) \leq \epsilon = 3\epsilon/3$.
- Não tipicidade possui probabilidade menor do que ϵ, ou seja, $\Pr(A_\epsilon^{(n)c}) \leq \epsilon$, implicando em $\Pr(A_\epsilon^{(n)}) \geq 1 - \epsilon$.
- A probabilidade de erro é menor do que ϵ.

□

Tipicidade Conjunta XI

$|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.

Temos a seguinte relação

$$1 = \sum_{(x^n,y^n)} p(x^n,y^n) \geq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n,y^n) \geq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)+\epsilon)}, \tag{620}$$

logo, teremos que

$$|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)} \tag{621}$$

...

Tipicidade Conjunta XII

$|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$.

continuação...

Como visto anteriormente, $\Pr(A_\epsilon^{(n)}) \geq 1-\epsilon$, para n grande suficiente, e assim teremos

$$1-\epsilon \leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} p(x^n,y^n) \leq |A_\epsilon^{(n)}| 2^{-n(H(X,Y)-\epsilon)} \tag{622}$$

e assim

$$|A_\epsilon^{(n)}| \leq (1-\epsilon)2^{n(H(X,Y)-\epsilon)} \tag{623}$$

□

Tipicidade Conjunta XIII

Duas seq. indep. provavelmente não são conjuntamente típica.

Seja $\tilde{X}^n$, $\tilde{Y}^n$ independente $\sim p(x^n)p(y^n)$, i.e., as duas sequências são independentes entre si. Vamos utilizar

$$A, B \perp\!\!\!\perp C \Rightarrow A \perp\!\!\!\perp C \text{ e } B \perp\!\!\!\perp C \tag{624}$$

e desta forma, teremos

$$\tilde{X}_i^n \perp\!\!\!\perp \tilde{Y}_j^n, \quad \forall i, j \tag{625}$$

...

Tipicidade Conjunta XIV

Duas seq. indep. provavelmente não são conjuntamente típica.

continuação...

Temos a seguinte derivação

$$\begin{aligned}
\Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}\right) &= \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} \underbrace{p(x^n)}_{\leq 2^{-n(H(X)-\epsilon)}} \underbrace{p(y^n)}_{\leq 2^{-n(H(Y)-\epsilon)}} \\
&\leq \sum_{(x^n,y^n)\in A_\epsilon^{(n)}} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\
&\leq 2^{n(H(X,Y)+\epsilon)} 2^{-n(H(X)-\epsilon)} 2^{-n(H(Y)-\epsilon)} \\
&= 2^{-n(I(X;Y)-3\epsilon)} \qquad (626)
\end{aligned}$$

□

Teorema da Codificação de Shannon I

Recordando o que foi visto...

- Existem $\approx 2^{nH(X)}$ sequências típicas em X.
- Existem $\approx 2^{nH(Y)}$ sequências típicas em Y.
- O total de pares independentes é $\approx 2^{nH(X)}2^{nH(Y)}$, mas nem todos eles são conjuntamente típicos. Apenas $\approx 2^{nH(X,Y)}$ deles são conjuntamente típicos.
- A proporção das sequências típicas independentes que são conjuntamente típicas é

$$\frac{2^{nH(X,Y)}}{2^{nH(X)}2^{nH(Y)}} = 2^{n(H(X,Y)-H(X)-H(Y))} = 2^{-nI(X;Y)} \tag{627}$$

  e isto representa a probabilidade de que um par de sequencias marginalmente típicas, escolhido aleatoriamente, seja conjuntamente típico.
- Se utilizarmos a tipicidade para decodificar então existem aproximadamente $2^{nI(X;Y)}$ pares de sequências disponíveis antes de precisarmos utilizar pares que seriam conjuntamente típicos se escolhidos aleatoriamente.
- Exemplo análogo: se $p(x) = 1/M$, então podemos escolher em média M amostras antes de vermos um determinado x em particular.

Teorema da Codificação de Shannon II

- Ideia básica é utilizar a tipicidade.
- Dada uma palavra recebida y^n, encontre um x^n que seja conjuntamente típico com y^n.
- x^n irá ocorrer conjuntamente com y^n com probabilidade ≈ 1, para n grande suficiente.
- Além disso, a probabilidade de que algum outro $\hat{x}^n$ seja conjuntamente típico com y^n será pequena, $\approx 2^{-nI(X;Y)}$.
- Se utilizarmos menos que $2^{nI(X;Y)}$ palavras, então a probabilidade de que alguma outra sequencia seja conjuntamente típica ocorrerá com probabilidade exponencialmente decrescente, para n grande suficiente.

Teorema (Teorema de Codificação de Shannon)

Todas as taxas menores do que $C \triangleq \max_{p(x)} I(X;Y)$ *são alcançáveis. Especificamente,* $\forall R < C$, *existe uma sequencia de códigos* $(2^{nR}, n)$ *com probabilidade máxima de erro* $\lambda^{(n)} \to 0$ *quando* $n \to \infty$. *Por outro lado, qualquer sequencia de códigos* $(2^{nR}, n)$ *com* $\lambda^{(n)} \to 0$ *quando* $n \to \infty$ *deverá ter* $R < C$.

Teorema da Codificação de Shannon III

- Enquanto não codificarmos acima da capacidade, poderemos codificar com erro nulo.
- Isto é verdadeiro para todo canal que possa ser representado por este modelo.

Teorema da Codificação de Shannon IV

- Podemos olhar para o erro de um código em particular e encontrar os limites deste erro.
- Ao invés, iremos verificar a probabilidade média de erro para todos códigos gerados de forma aleatória.
- Iremos mostrar que este erro médio é pequeno.
- Isto implica que $\exists$ muitos códigos bons para que seja possível fazer a média ser pequena.
- Para mostrar que a probabilidade máxima de erro também é pequena, iremos descartar 50% dos códigos.
- Ideia: para um dado canal $(\mathcal{X}, p(y|x), \mathcal{Y})$ vamos tomar um código $(2^{nR}, n)$ com taxa R. Isto significa que precisaremos de:
  1) Conjunto de índices: $\{1, \ldots, M\}$;
  2) Codificador: $X^n : \{1, \ldots, M\} \to \mathcal{X}^n$, mapeando na palavra $X^n(i)$;
  3) Decodificador: $g : \mathcal{Y}^n \to \{1, \ldots, M\}$.
- A demonstração terá duas partes: 1) mostrar que todas as taxas $R < C$ são alcançáveis (existe um código com probabilidade de erro evanescente); 2) se o erro tende a zero, devemos ter $R < C$.

Teorema da Codificação de Shannon V

todas as taxas $R < C$ são alcançáveis.

- Dado $R < C$, vamos assumir que utilizaremos uma distribuição $p(x)$ arbitrária e geraremos 2^{nR} palavras aleatórias utilizando a extensão $p(x^n) = \prod_{i=1}^{n} p(x_i)$.
- O conjunto de palavras, nosso *codebook*, pode ser representado pela matriz

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & x_3(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & x_3(2) & \dots & x_n(2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & x_3(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix} \tag{628}$$

Desta forma, existe 2^{nR} palavras (*codewords*), de comprimento n, geradas por $p(x)$, cada uma em uma linha da matriz.

...

Teorema da Codificação de Shannon VI

todas as taxas $R < C$ são alcançáveis.

continuação...

- O *codebook* $\mathcal{C}$ é aleatório e terá probabilidade $\Pr(\mathcal{C})$.
- Para enviar qualquer mensagem $w \in \{1, \ldots, M = 2^{nR}\}$, devemos enviar a palavra correspondente $x_{1:n}(w) = (x_1(w), x_2(w), \ldots, x_n(w))$.
- Podemos calcular a probabilidade de uma dada palavra $w \in \{1, \ldots, M\}$ da seguinte forma

$$p\left(x^n(w)\right) = \prod_{i=1}^{n} p\left(x_i(w)\right), \quad w \in \{1, \ldots, M\}. \tag{629}$$

...

Teorema da Codificação de Shannon VII

todas as taxas $R < C$ são alcançáveis.

continuação...

- A probabilidade de todo o *codebook* será dada por

$$p(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p\left(x_i(w)\right). \tag{630}$$

...

Teorema da Codificação de Shannon VIII

todas as taxas $R < C$ são alcançáveis.

continuação...

Considere o seguinte esquema de codificação/decodificação:

1) Gere um *codebook* aleatório, como descrito anteriormente, de acordo com uma distribuição $p(x)$;
2) O *codebook* é conhecido pelo transmissor/receptor (também conhecem $p(y|x)$).
3) Vamos gerar mensagens W de acordo com uma distribuição uniforme, ou seja, todas mensagens serão equiprováveis, $p(W = w) = 2^{-nR}$ para $w = 1, \ldots, 2^{nR}$.
4) Escolhemos uma mensagem w e enviamos a palavra correspondente $x^n(w)$ através do canal.

...

Teorema da Codificação de Shannon IX

todas as taxas $R < C$ são alcançáveis.

continuação...

5) A mensagem recebida Y^n será dada de acordo com a seguinte distribuição

$$Y^n \sim p(y^n \mid x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)) \tag{631}$$

6) O sinal é decodificado utilizando a decodificação através do conjunto típico (será descrito adiante).

...

Teorema da Codificação de Shannon X

todas as taxas $R < C$ são alcançáveis.

continuação...

Decodificador utilizando conjunto típico. Vamos decodificar $\hat{w}$ se

1) $(x^n(\hat{w}), y^n)$ é conjuntamente tipico, ou seja, $(x^n(\hat{w}), y^n) \in A_\epsilon^{(n)}$;
2) $\hat{w}$ é único, ou seja, $\nexists w'$ tal que $(x^n(w'), y^n) \in A_\epsilon^{(n)}$ para $w' \neq \hat{w}$.

Se estas duas condições não acontecerem, teremos um erro e a saída será um inteiro especial '0' (para designar erro). Três tipos de erro podem ocorrer:

a) $\exists w' \neq \hat{w}$ tal que $(x^n(w'), y^n) \in A_\epsilon^{(n)}$, i.e., existe mais do que uma mensagem típica;
b) $\nexists \hat{w}$ tal que $(x^n(\hat{w}), y^n)$ é conjuntamente tipico;
c) se $\hat{w} \neq w$, i.e., a palavra errada é conjuntamente típica.

...

Teorema da Codificação de Shannon XI

todas as taxas $R < C$ são alcançáveis.

continuação...

obs.: decodificação de máxima verossimilhança é ótima, já a decodificação pelo conjunto típico não é ótima, entretanto será suficiente para mostrar o resultado pretendido.

...

Teorema da Codificação de Shannon XII

todas as taxas $R < C$ são alcançáveis.

continuação...

Vejamos três medidas de qualidade que poderemos utilizar.

1) Erro específico do código:

$$P_e^{(n)}(\mathcal{C}) = \Pr(\hat{w} \neq w \mid \mathcal{C}) = \frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} \lambda_i \tag{632}$$

onde (conforme definido anteriormente)

$$\lambda_i \triangleq \Pr(g(Y^n)) \neq i \mid X^n = X^n(i)) = \sum_{y^n \in \mathcal{Y}^n} p(y^n \mid X^n(i)) \mathbf{1}_{\{g(y^n) \neq i\}}. \tag{633}$$

Note que $\mathcal{C}$ é uma v.a..

...

Teorema da Codificação de Shannon XIII

todas as taxas $R < C$ são alcançáveis.

continuação...

2) Erro médio entre todos códigos gerados aleatoriamente (vamos tomar o valor esperado do erro específico de um código)

$$\begin{aligned} \Pr(\varepsilon) &= \mathrm{E}[P_e^n(\mathcal{C})] \\ &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \Pr(\hat{W} \neq W | \mathcal{C}) \\ &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \end{aligned} \tag{634}$$

Isto será mais fácil de analisar do que P_e.

...

Teorema da Codificação de Shannon XIV

todas as taxas $R < C$ são alcançáveis.

continuação...

3) Erro máximo de um código:

$$P_{\mathcal{C},\max}(\mathcal{C}) = \max_{i \in \{1,2,\ldots,M\}} \lambda_i. \tag{635}$$

Queremos mostrar que se $R < C$, então existe um *codebook* $\mathcal{C}$ tal que $P_{\mathcal{C},\max} \to 0$ (e se $R > C$, então $P_{\mathcal{C},\max} \to 1$).

...

Teorema da Codificação de Shannon XV

todas as taxas $R < C$ são alcançáveis.

continuação...

Procedimento:

1) Expandir o erro médio e mostrar que ele é pequeno;
2) Deduzir que $\exists$ ao menos 1 código com erro pequeno;
3) Mostrar que isto pode ser modificado para termos uma probabilidade de erro máximo pequena.

...

Teorema da Codificação de Shannon XVI

todas as taxas $R < C$ são alcançáveis.

continuação...

Erro médio

$$\begin{aligned}\Pr(\varepsilon) &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C}) \\ &\quad \text{valor médio, entre os codebooks, dos} \\ &\quad \text{valores médios entre as palavras em um coodebook} \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}) \qquad (636)\end{aligned}$$

...

Teorema da Codificação de Shannon XVII

todas as taxas $R < C$ são alcançáveis.

continuação...

mas

$$\sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C}) =$$

$$\sum_{\mathcal{C}} \overbrace{\Pr\left(g(Y^n) \neq w \mid X^n = x^n(w)\right)}^{\lambda_w(\mathcal{C})} \overbrace{\Pr\left(x^n(1), \ldots, x^n(2^{nR})\right)}^{\Pr(\mathcal{C})} =$$

$$\sum_{\mathcal{C}} \underbrace{\Pr\left(g(Y^n) \neq w \mid X^n = x^n(w)\right) \overbrace{\Pr\left(x^n(1), \ldots, x^n(2^{nR})\right)}^{\prod_{i=1}^{2^{nR}} \Pr(x^n(i))}}_{\text{coisa}} =$$

$$\sum_{x^n(1),\ldots,x^n(2^{nR})} \text{coisa} \qquad (637)$$

...

Teorema da Codificação de Shannon XVIII

todas as taxas $R < C$ são alcançáveis.

continuação...

$$\sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C}) = \sum_{x^n(1),\ldots,x^n(2^{nR})} \text{coisa}$$

$$\text{somatório com } \underbrace{2^n \cdots 2^n}_{2^{nR}\text{vezes}} = 2^{n2^R} \text{termos}$$

$$= \sum_{\substack{x^n(1),\ldots,x^n(w-1),\\ x^n(w+1),\ldots,x^n(2^{nR})}} \sum_{x^n(w)} \text{coisa} \qquad (638)$$

...

Teorema da Codificação de Shannon XIX

todas as taxas $R < C$ são alcançáveis.

continuação...

Temos também que

$$\begin{aligned} \text{coisa} &= \prod_{i=1}^{2^{nR}} \Pr(x^n(i))\lambda_w(\mathcal{C}) \\ &= \left(\prod_{\substack{i=1 \\ i \neq w}}^{2^{nR}} \Pr(x^n(i)) \right) \underbrace{\Pr(x^n(w))\lambda_w(\mathcal{C})}_{\text{termo em } w} \end{aligned} \tag{639}$$

...

Teorema da Codificação de Shannon XX

todas as taxas $R < C$ são alcançáveis.

continuação...

$$\sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C}) = \underbrace{\sum_{\substack{x^n(1),\ldots,x^n(w-1),\\ x^n(w+1),\ldots,x^n(2^{nR})}} \underbrace{p\begin{pmatrix} x^n(1),\ldots,x^n(w-1),\\ x^n(w+1),\ldots,x^n(2^{nR}) \end{pmatrix}}_{\prod_{i\neq w}\Pr(x^n(i))} \sum_{x^n(w)} \frac{\Pr(g(Y^n)\neq w | X^n = x^n(w))}{\Pr(x^n(w))}}_{=1} = \quad (640)$$

O primeiro termo é igual à 1 pois corresponde à soma dos termos de uma distribuição conjunta ($2^{nR} - 1$ termos).

...

Teorema da Codificação de Shannon XXI

todas as taxas $R < C$ são alcançáveis.

continuação...

$$\sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C}) = \sum_{x^n(w)} \Pr\left(g(Y^n) \neq w \mid X^n = x^n(w)\right) \Pr\left(x^n(w)\right) = \sum_{x^n \in \mathcal{X}^n} \Pr\left(g(Y^n) \neq 1 \mid X^n = x^n(1)\right) \Pr\left(x^n(1)\right) \quad (641)$$

...

Teorema da Codificação de Shannon XXII

todas as taxas $R < C$ são alcançáveis.

continuação...

Na equação anterior utilizamos que o resultado é o mesmo para qualquer w, logo, sem perca de generalidade, escolhemos $w = 1$, uma mensagem qualquer. Teremos assim

$$\sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_w(\mathcal{C}) = \sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_1(\mathcal{C}) = \beta \tag{642}$$

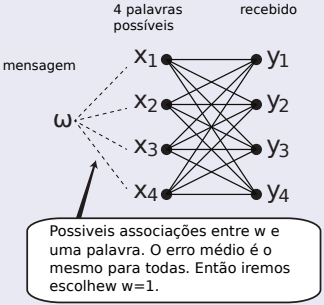
...

Teorema da Codificação de Shannon XXIII

todas as taxas $R < C$ são alcançáveis.

continuação...

Exemplo: intuição sobre como chegamos a β.



...

Teorema da Codificação de Shannon XXIV

todas as taxas $R < C$ são alcançáveis.

continuação...

O erro então é igual a: prob. de escolher x_1 para w e não escolher y_1 + prob. de escolher x_2 para w e não escolher y_2 + ...

Teremos o mesmo resultado para todo $w \in \{1, 2, \ldots, M\}$, então podemos selecionar $w = 1$.

Obtemos assim

$$\Pr(\varepsilon) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \beta = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) = \beta \tag{643}$$

onde $\beta = \Pr(\varepsilon \mid W = 1)$.

...

Teorema da Codificação de Shannon XXV

todas as taxas $R < C$ são alcançáveis.

continuação...

Vamos definir eventos de erro aleatórios (considerando $w = 1$)

$$E_i \triangleq \left\{(x^n(i), y^n) \in A_\epsilon^{(n)}\right\} \text{ para } i = 1, \ldots, 2^{nR} \tag{644}$$

Assumindo que a entrada é $x^n(1)$, então não haver erro é equivalente a

$$E_1 \cap \neg(E_2 \cup E_3 \cup \ldots \cup E_M). \tag{645}$$

...

Teorema da Codificação de Shannon XXVI

todas as taxas $R < C$ são alcançáveis.

continuação...

Tipos de erro:

- E_1^c significa que a palavra transmitida e recebida não são conjuntamente típicas (erro tipo B).
- $E_2 \cup E_3 \cup \ldots \cup E_{2^{nR}}$ é uma das seguintes possibilidades:
  - erro tipo C: a palavra errada é conjuntamente típica com a sequência recebida;
  - erro tipo A: mais do que 1 palavra é conjuntamente típica com a sequência recebida.

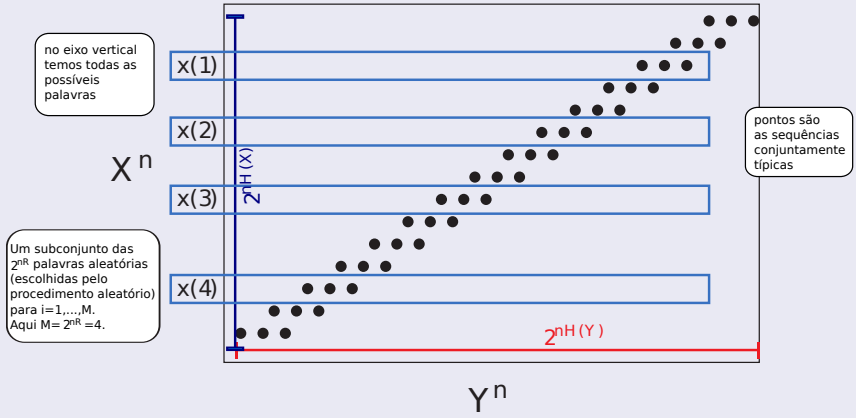
...

Teorema da Codificação de Shannon XXVII

todas as taxas $R < C$ são alcançáveis.

Teorema da Codificação de Shannon XXVIII

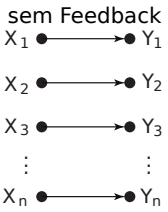
continuação...

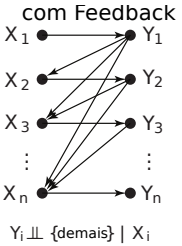


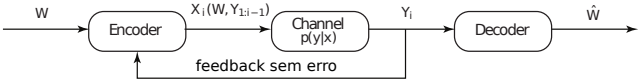
...

Feedback I

Em um canal discreto sem memória, se adicionarmos *feedback* a este canal de comunicação, teremos uma taxa R maior? Mostraremos que, neste caso, a taxa não aumenta ao utilizar *feedback*.







Feedback II

- O feedback pode tornar a decodificação mais simples.
- O feedback pode ajudar quando temos um canal com memória.
- No caso de um canal sem memória, feedback não ajuda a melhorar a taxa.

Feedback III

Definição (código $(2^{nR}, n)$ com *feedback*)

Um código desta forma é dado por um codificador $X_i(W, Y_{1:i-1})$, um decodificador $g : Y^n \to \{1, 2, \ldots, 2^{nR}\}$ e $P_e^{(n)} = \Pr(g(Y^n) \neq W)$ para $H(W) = nR$ (uniforme).

Definição (Capacidade)

A capacidade de um canal discreto sem memória com *feedback* (C_{fb}) é o máximo de todas as taxas alcançáveis por códigos com *feedback*.

Feedback IV

Teorema

$$C_{fb} = C = \max_{p(x)} I(X;Y) \tag{674}$$

para um canal discreto sem memória.

Demonstração.

- $C_{\text{fb}} \geq C$, já que *feedback* é uma generalização.
- Vamos utilizar W ao invés de X e limitar R.

$$\begin{aligned} H(W) &= H(W \mid Y^n) + I(W;Y^n) \\ &\leq 1 + P_e^{(n)} nR + I(W;Y^n) \quad \text{Fano} \end{aligned} \tag{675}$$

- Vamos limitar $I(W;Y^n)$.

...

Feedback V

Demonstração.

continuação...

$$
\begin{aligned}
I(W;Y^n) &= H(Y^n) - H(Y^n \mid W) \quad \text{definição} \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i \mid Y_{1:i-1}, W) \quad \text{regra da cadeia} \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i \mid Y_{1:i-1}, W, X_i) \quad \text{pois } X_i = f(W, Y_{1:i-1}) \\
&\quad Y_i \perp\!\!\!\perp Y_{1:i-1}, W \mid X_i \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i \mid X_i) \qquad (676)
\end{aligned}
$$

...

Feedback VI

Demonstração.

continuação...

$$
\begin{aligned}
I(W;Y^n) &= H(Y^n) - \sum_{i=1}^{n} H(Y_i \mid X_i) \\
&\leq \sum_{i} H(Y_i) - \sum_{i=1}^{n} H(Y_i \mid X_i) \\
&\leq \sum_{i} I(X_i;Y_i) \leq nC \qquad (677)
\end{aligned}
$$

...

Feedback VII

Demonstração.

continuação...

Teremos então

$$H(W) \leq 1 + P_e^{(n)} nR + nC \tag{678}$$

e assim teremos

$$nR \leq 1 + P_e^{(n)} nR + nC \tag{679}$$

pois a primeira desigualdade é válida para todo $H(W)$, inclusive para o máximo, quando $H(W) = nR$. Desta forma

$$R \leq \frac{1}{n} + P_e^{(n)} R + C \tag{680}$$

ou $R \leq C$ quando $n \to \infty$. Então o *feedback* não traz benefício.

□

Teorema Conjunto da Fonte/Canal I

- Compressão de dados: é possível comprimir sem erro a uma taxa média $R > H$ (bits por símbolo da fonte), onde H é a entropia da fonte.
- Transmissão de dados: é possível realizar uma comunicação sem erro se a taxa média é tal que $R < C$ (bits por utilização do canal), onde C é a capacidade de canal.
- Parece intuitivo então que seremos capazes de enviar informação de forma confiável de uma fonte com entropia H, através de um canal com capacidade C, se $H < C$.
- O canal pode ser utilizado para transmitir dados de fontes diferentes.
- Separação entre (de)codificador de fonte e (de)codificador de canal.

Teorema Conjunto da Fonte/Canal II

- Fonte: $V \in \mathcal{V}$ que satisfaz a propriedade da equipartição assintótica.
- Enviar $V_{1:n} = V_1, V_2, \ldots, V_n$ através de um canal.
- Informação por símbolo, taxa de entropia, $H(\mathcal{V})$ do processo estocástico (se i.i.d., teremos $H(\mathcal{V}) = H(V_i), \forall i$).
- $V_{1:n} \rightarrow$ codificador $\rightarrow X^n \rightarrow$ canal $\rightarrow Y^n$decodificador $\rightarrow \hat{V}_{1:n}$
- probabilidade de erro:

$$\begin{aligned} P_e^{(n)} &= P(V_{1:n} \neq \hat{V}_{1:n}) \\ &= \sum_{y_{1:n}} \sum_{v_{1:n}} \Pr(v_{1:n}) \Pr(y_{1:n} \mid X^n(v_{1:n})) \mathbf{1}_{\{g(y_{1:n}) \neq v_{1:n}\}} \end{aligned} \tag{681}$$

Teorema Conjunto da Fonte/Canal III

Teorema (Teorema da Codificação de Fonte/Canal)

Se $V_{1:n}$ é um processo estocástico com alfabeto finito que satisfaz a prop. da eq. ass. e $H(\mathcal{V}) < C$, então $\exists$ uma sequência de códigos $(2^{nR}, n)$ com $P_e^{(n)} \to 0$. Por outro lado, se $H(\mathcal{V}) > C$, então $P_e^{(n)} > 0$ para qualquer n e não será possível enviar informação com probabilidade de erro arbitrariamente pequena.

Teorema Conjunto da Fonte/Canal IV

Demonstração.

- Se V satisfaz a prob. da eq. ass., então $\exists$ um conjunto $A_\epsilon^{(n)}$ com $|A_\epsilon^{(n)}| \leq 2^{n(H(\mathcal{V})+\epsilon)}$, e neste conjunto está contido 'tudo' aquilo que ocorre.
- Iremos codificar apenas o conjunto típico e enviar um erro caso contrário. Isto contribui em, no máximo, ϵ para a P_e.
- Vamos indexar os elementos de $A_\epsilon^{(n)}$ com $\{1, 2, \ldots, 2^{n(H+\epsilon)}\}$, precisaremos então de $n(H+\epsilon)$ bits.
- A taxa será $R = H(\mathcal{V}) + \epsilon$. Se $R < C$ então a probabilidade de erro será menor que ϵ, e poderemos fazê-la tão pequena quanto desejado.

...

Teorema Conjunto da Fonte/Canal V

Demonstração.

continuação...

$$\begin{aligned} P_e^{(n)} &= \Pr(V_{1:n} \neq \hat{V}_{1:n}) \\ &\leq \Pr(V_{1:n} \notin A_\epsilon^{(n)}) + \underbrace{\Pr(g(Y^n) \neq V^n \mid V^n \in A_\epsilon^{(n)})}_{<\epsilon, \text{ já que } R<C} \\ &\leq \epsilon + \epsilon = 2\epsilon \end{aligned} \tag{682}$$

É possível assim reconstruir a sequência com baixa probabilidade de erro para n grande suficiente, se

$$H(\mathcal{V}) < C. \tag{683}$$

□

Teorema Conjunto da Fonte/Canal VI

proposição inversa.

- Para mostrar que $P_e^{(n)} \to 0 \Rightarrow H(\mathcal{V}) \leq C$ iremos definir

$$\text{codificador: } X^n(V^n) : \mathcal{V}^n \to \mathcal{X}^n \tag{684}$$

$$\text{decodificador: } g(Y^n) : \mathcal{Y}^n \to \mathcal{V}^n \tag{685}$$

- Fano: $H(X \mid Y) \leq 1 + P_e \log |\mathcal{X}|$.
- Teremos então:

$$H(V^n \mid \hat{V}^n) \leq 1 + P_e^{(n)} \log |\mathcal{V}^n| = 1 + nP_e \log |\mathcal{V}| \tag{686}$$

...

Teorema Conjunto da Fonte/Canal VII

proposição inversa.

continuação...

$$
\begin{aligned}
H(\mathcal{V}) &\leq \frac{H(V_1, V_2, \ldots, V_n)}{n} = \frac{H(V_{1:n})}{n} \\
&= \frac{1}{n} H(V_{1:n} \mid \hat{V}_{1:n}) + \frac{1}{n} I(V_{1:n}; \hat{V}_{1:n}) \\
&\quad \text{Fano} \\
&\leq \frac{1}{n}(1 + P_e^{(n)} n \log |\mathcal{V}|) + \frac{1}{n} I(V_{1:n}; \hat{V}_{1:n}) \\
&\quad \text{desig. proc. dados } V \to X \to Y \to \hat{V} \\
&\leq \frac{1}{n}(1 + P_e^{(n)} n \log |\mathcal{V}|) + \frac{1}{n} I(X_{1:n}; Y_{1:n}) \qquad (687)
\end{aligned}
$$

...

Teorema Conjunto da Fonte/Canal VIII

proposição inversa.

continuação...

$$
\begin{aligned}
H(\mathcal{V}) &\leq \frac{1}{n}(1 + P_e^{(n)} n \log |\mathcal{V}|) + \frac{1}{n} I(X_{1:n}; Y_{1:n}) \\
&\quad \text{otimização capacidade, sem memória} \\
&\leq \frac{1}{n} + P_e^{(n)} \log |\mathcal{V}| + C \qquad (688)
\end{aligned}
$$

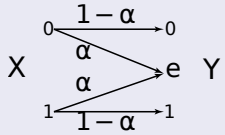
Fazendo $n \to \infty$, teremos $1/n \to 0$ e $P_e \to 0$, assim obtemos $H(\mathcal{V}) \leq C$. □

Canal Binário com Apagamento I

3 formas de calcular a capacidade do canal binário com apagamento

Exercício (Capacidade do canal binário com apagamento)

Considere o canal binário com apagamento ilustrado abaixo.



A capacidade do canal é dada por:

$$C = \max_{p(x)} I(X;Y) = \max_{p(x)} H(X) - H(X|Y) = \max_{p(x)} H(Y) - H(Y|X) \tag{689}$$

Podemos adotar dois caminhos, usando $H(X) - H(X|Y)$ *ou* $H(Y) - H(Y|X)$.

...

Canal Binário com Apagamento II

Exercício (Capacidade do canal binário com apagamento)

continuação...

Calcular $H(X)$ depende apenas de $p(x)$, teremos $H(X) = H(p)$.
Para calcular $H(X|Y)$ devemos analisar cada uma das possíveis saídas:

$$
\begin{aligned}
H(X|Y) &= \sum_y p(y) H(X|Y=y) && (690)\\
&= p(Y=0)\underbrace{H(X|Y=0)}_{=0} + \underbrace{p(Y=e)}_{=\alpha}\underbrace{H(X|Y=e)}_{=H(p)} + \\
&\quad p(Y=1)\underbrace{H(X|Y=1)}_{=0} \\
&= \alpha H(p) && (691)
\end{aligned}
$$

...

Canal Binário com Apagamento III

Exercício (Capacidade do canal binário com apagamento)

continuação...

Logo, teremos $I(X;Y) = H(p) - \alpha H(p) = (1-\alpha)H(p)$.

$$C = \max_{p}(1-\alpha)H(p) = (1-\alpha), \tag{692}$$

onde a distribuição p *que maximiza é a uniforme* $p = (0.5, 0.5)$.

...

Canal Binário com Apagamento IV

Exercício (Capacidade do canal binário com apagamento)

continuação...

Para o segundo caminho deveremos calcular $H(Y)$ e $H(Y|X)$. O cálculo de $H(Y|X)$ é simples:

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ &= p_0 \underbrace{H(Y|X=0)}_{=H(\alpha)} + p_1 \underbrace{H(Y|X=1)}_{=H(\alpha)} \\ &= H(\alpha). \end{aligned} \tag{693}$$

...

Canal Binário com Apagamento V

Exercício (Capacidade do canal binário com apagamento)

continuação...

Para calcular $H(Y)$ podemos calcular as probabilidades de $Y = 0$, $Y = e$ e $Y = 1$, e então calcular

$$\begin{aligned} H(Y) &= H(\Pr(Y=0), \Pr(Y=e), \Pr(Y=1)), \\ &= H\left(p_0(1-\alpha), \alpha, p_1(1-\alpha)\right) \end{aligned} \tag{694}$$

*conforme feito anteriormente (ver slide **??**).*

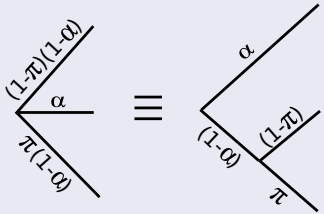
...

Canal Binário com Apagamento VI

Exercício (Capacidade do canal binário com apagamento)

continuação...

Ou ainda, podemos ver que a determinação de Y pode ser quebrada em duas etapas: 1) houve apagamento ou não (entropia associada a esta etapa é $H(\alpha)$); 2) caso não tenha ocorrido apagamento (o que ocorre com probabilidade $1-\alpha$), Y será 0 ou 1 (entropia associada a esta etapa é $H(p)$.



...

Canal Binário com Apagamento VII

Exercício (Capacidade do canal binário com apagamento)

continuação...

Teremos assim que

$$H(Y) = H(\alpha) + (1-\alpha)H(p). \tag{695}$$

Desta forma, teremos

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max_{p(x)} H(Y) - H(Y|X) \\ &= \max_{p(x)} H(\alpha) + (1-\alpha)H(p) - H(\alpha) \\ &= \max_{p(x)} (1-\alpha)H(p) = 1-\alpha, \end{aligned} \tag{696}$$

onde a distribuição p que maximiza é a uniforme $p = (0.5, 0.5)$.

Capacidade de Canal I

Exercício (Capacidade de Canal)

Considere um canal discreto em que os alfabetos de entrada e saída são $\mathcal{X} = \mathcal{Y} = \{0, 1, 2\}$. *As probabilidades de transição para este canal são dadas:* $p(y = 0|x = 0) = p(y = 2|x = 2) = 1$, $p(y = 0|x = 1) = p(y = 2|x = 1) = 1/4$ *e* $p(y = 1|x = 1) = 1/2$.

...

Capacidade de Canal II

Exercício (Capacidade de Canal)

continuação...

a) *Ache a informação mútua entre a entrada do canal e a saída se as probabilidades da entrada são dadas por* $p(x=0)=1/2$, $p(x=1)=0$ *e* $p(x=2)=1/2$.
(solução)

$$I(X;Y) = \underbrace{H(X)}_{=1} - \underbrace{H(X|Y)}_{=0} = 1 \tag{697}$$

onde utilizamos que, dado Y, *como* $y=1$ *não ocorre, então não existe incerteza sobre* X, *uma vez que sendo enviado* 0 *ou* 2, *iremos receber o mesmo símbolo.*

...

Capacidade de Canal III

Exercício (Capacidade de Canal)

continuação...

b) *Encontre a informação mútua entre a saída e entrada no canal quando a probabilidade da entrada é uniforme.*

(solução)
$I(X;Y) = H(Y) - H(Y|X)$. *Vamos calcular então as probabilidades das saídas*
$p(y=0) = p(x=0) + (1/4)p(x=1) = 5/12$; $p(y=1) = (1/2)p(x=1) = 1/6$; *e*
$p(y=2) = p(x=2) + (1/4)p(x=1) = 5/12$. ...

Capacidade de Canal IV

Exercício (Capacidade de Canal)

continuação...

$$H(Y) = \frac{5}{12}\log\frac{12}{5} + \frac{1}{6}\log 6 + \frac{5}{12}\log\frac{12}{5} = 1.483 \tag{698}$$

$$\begin{aligned} H(Y|X) &= \frac{1}{3}\underbrace{H(Y|X=0)}_{=0} + \frac{1}{3}H(Y|X=1) + \frac{1}{3}\underbrace{H(Y|X=2)}_{=0} \\ &= \frac{1}{3}H(Y|X=1) \\ &= \frac{1}{3}\left[\frac{1}{4}\log 4 + \frac{1}{2}\log 2 + \frac{1}{4}\log 4\right] = \frac{1}{3}\times\frac{3}{2} = \frac{1}{2} \end{aligned} \tag{699}$$

...

Capacidade de Canal V

Exercício (Capacidade de Canal)

continuação...

e assim a informação é dada por

$$I(X;Y) = H(Y) - H(Y|X) = 1.483 - 0.5 = 0.983 \text{ bits.} \quad (700)$$

...

Capacidade de Canal VI

Exercício (Capacidade de Canal)

continuação...

c) *Encontre a capacidade destes canal e uma distribuição para a entrada capaz de atingir esta capacidade.*

(solução)
Queremos maximizar $I(X;Y)$ *com relação a* p_0, p_1, p_2 ($p(x=0), p(x=1), p(x=2)$, *respectivamente), que devem ser positivos e somar* 1. *Vamos chamar de* $q_0 = p(y=0)$, $q_1 = p(y=1)$ *e* $q_2 = p(y=2)$.
Temos então: $q_0 = p_0 + (1/4)p_1$, $q_1 = (1/2)p_1$ *e* $q_2 = p_2 + (1/4)p_1$.
Iremos mostrar que $I(X;Y)$ *poderá ser maximizada quando a distribuição de entrada for tal que* $p_0 = p_2$.

...

Capacidade de Canal VII

Exercício (Capacidade de Canal)

continuação...

Podemos escrever $p_2 = 1 - p_1 - p_0$.
As probabilidades de saída serão $q_0 = p_0 + (1/4)p_1$, $q_1 = (1/2)p_1$ *e* $q_2 = p_2 + (1/4)p_1$.
$I(X;Y) = H(Y) - H(Y|X)$. *Note que* $H(Y|X) = p_1 H(Y|X = 1)$ *e, desta forma, depende apenas de* p_1. *Dado um valor para* p_1, *teremos* q_1 *determinado e assim* $H(Y)$ *será maximizado quando* q_0 *e* q_2 *forem iguais, o que irá ocorrer quando* $p_0 = p_2$. *Logo, fixando* p_1, *o máximo de* $I(X;Y)$ *será alcançado quando* $p_0 = p_2 = (1 - p_1)/2$. *Para este valor de* p_0 *e* p_2, *teremos* $q_0 = q_2 = (2 - p_1)/4$.

...

Capacidade de Canal VIII

Exercício (Capacidade de Canal)

continuação...

Teremos assim

$$\begin{aligned} H(Y) &= -q_0 \log q_0 - q_1 \log q_1 - q_2 \log q_2 \\ &= -2\left((2-p_1)/4 \log(2-p_1)/4\right) - (p_1/2)\log(p_1/2) \end{aligned} \tag{701}$$

A informação mútua para $p_0 = p_2$ será dada por

$$I(X;Y) = -(1-p_1/2)\log((2-p_1)/4) - (p_1/2)\log(p_1/2) - (3/2)p_1 \tag{702}$$

...

Capacidade de Canal IX

Exercício (Capacidade de Canal)

continuação...

```
p1 = [0.001 : 0.001 : 0.999];
I = -(1-p1/2).*log2((2-p1)/4) - (p1/2).*log2(p1/2) - (3/2)*p1;
plot(p1,I); xlabel('p1'); ylabel('I(X;Y)');
[ml, id] = max(I);
p1(id)
ans =  0.11800
ml
ml =  1.0875
```

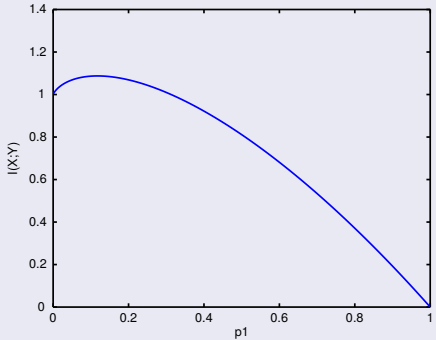
O máximo ocorrerá quando $p_1 = 0.118$ *e* $p_0 = p_2 = 0.441$.

...

Capacidade de Canal X

Exercício (Capacidade de Canal)

continuação...



Canal da Soma I

Exercício (Canal da Soma)

Seja $\mathcal{X} = \mathcal{Y} = \{A, B, C, D\}$ os alfabetos de entrada e saída de um canal discreto sem memória com probabilidades de transição $p(y|x)$ dadas pela matriz abaixo, para $0 \leq \epsilon, \delta \leq 1$,

$$p(y|x) = \begin{pmatrix} 1-\epsilon & \epsilon & 0 & 0 \\ \epsilon & 1-\epsilon & 0 & 0 \\ 0 & 0 & 1-\delta & \delta \\ 0 & 0 & \delta & 1-\delta \end{pmatrix} \tag{703}$$

...

Canal da Soma II

Exercício (Canal da Soma)

continuação...

Note que este canal com 4 entradas e saídas equivale à soma ou união de dois canais em paralelo com matrizes de transição dadas por

$$p_1(y|x) = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \quad (704) \qquad p_2(y|x) = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix} \quad (705)$$

com alfabetos $\mathcal{X}_1 = \mathcal{Y}_1 = \{A, B\}$

...

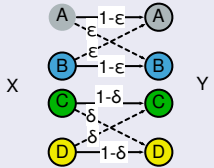
Canal da Soma III

Exercício (Canal da Soma)

continuação...

a) *Esboce o diagrama de transições deste canal.*

(solução)



...

Canal da Soma IV

Exercício (Canal da Soma)

continuação...

b) *Encontre a capacidade do canal para* $\epsilon = \delta = 1/2$.

(solução)
Nesta situação ($\epsilon = \delta = 1/2$) *temos uma canal simétrico, e sua capacidade é alcançada por uma distribuição uniforme. A capacidade será dada por*

$$\begin{aligned} C &= \log|\mathcal{Y}| - H(r) \qquad (706) \\ &\quad \textit{onde } r \textit{ é uma linha da matriz de transição} \\ &= \log 4 - H(\frac{1}{2}, \frac{1}{2}, 0, 0) \\ &= 2 - 1 = 1 \textit{ (bit por utilização do canal)} \end{aligned}$$

...

Canal da Soma V

Exercício (Canal da Soma)

continuação...

c) *Seja $p(x)$ uma função massa probabilidade em $\mathcal{X}$ e seja $p(A) + p(B) = \alpha$ e $p(C) + p(D) = 1 - \alpha$. Mostre que a informação mútua entre X e Y poderá ser expressa por*

$$I(X;Y) = H(\alpha) + \alpha I(X;Y|X \in \{A,B\}) + (1-\alpha) I(X;Y|X \in \{C,D\}). \tag{707}$$

(solução) Vamos definir uma v.a. $\theta \in \{0,1\}$ tal que, $x \in \{A,B\} \Rightarrow \theta = 1$ e $x \in \{C,D\} \Rightarrow \theta = 0$, logo a função massa de probabilidade de θ será dada por

$$p_\theta = \begin{cases} \alpha, & \theta = 1 \\ 1-\alpha, & \theta = 0 \end{cases} \tag{708}$$

...

Canal da Soma VI

Exercício (Canal da Soma)

continuação...

A informação mútua entre X e Y poderá ser expressa como

$$\begin{aligned} I(X;Y) &= I(X;\theta) + I(X;Y|\theta) \\ &= H(\theta) - H(\theta|X) + I(X;Y|\theta) \\ &= H(\alpha) - 0 + p(\theta=1)I(X;Y|\theta=1) + p(\theta=0)I(X;Y|\theta=0) \\ &= H(\alpha) + \alpha I(X;Y|X\in\{A,B\}) + (1-\alpha)I(X;Y|X\in\{C,D\}) \end{aligned} \quad (709)$$

...

Canal da Soma VII

Exercício (Canal da Soma)

continuação...

d) *Seja C_1 e C_2 as capacidades dos canais descritos por $p_1(y|x)$ e $p_2(y|x)$. Mostre que*

$$\max_{p(x)} I(X;Y) = \max_{\alpha} \left(H(\alpha) + \alpha C_1 + (1-\alpha)C_2\right) \tag{710}$$

(solução)
Note que

$$C_1 = \max_{p_1(x):x\in\{A,B\}} I(X;Y) \tag{711}$$

$$C_2 = \max_{p_2(x):x\in\{C,D\}} I(X;Y) \tag{712}$$

...

Canal da Soma VIII

Exercício (Canal da Soma)

Canal da Soma IX

continuação...

Como $x \in \{A, B\} \rightarrow y \in \{A, B\}$ *e* $x \in \{C, D\} \rightarrow y \in \{C, D\}$, *teremos*

$$\begin{aligned}
\max_{p(x):x\in\{A,B,C,D\}} I(X;Y) &= \\
\max_{p(x):x\in\{A,B,C,D\},0\leq\alpha\leq 1} (H(\alpha) + \alpha I(X;Y|X \in \{A,B\}) + & \\
(1-\alpha)I(X;Y|X \in \{C,D\})) &= \\
\max_{0\leq\alpha\leq 1} \left(H(\alpha) + \max_{p(x):x\in\{A,B\}} \alpha I(X;Y|X \in \{A,B\}) + \right. & \\
\left. \max_{p(x):x\in\{C,D\}} (1-\alpha) I(X;Y|X \in \{C,D\}) \right) &= \\
\max_{0\leq\alpha\leq 1} (H(\alpha) + \alpha C_1 + (1-\alpha) C_2) & \qquad (713)
\end{aligned}$$

...

Canal da Soma X

Exercício (Canal da Soma)

continuação...

e) *Encontre a capacidade C do canal de soma em termos das capacidades C_1 e C_2 dos sub-canais, sem utilizar outros parâmetros.*

(solução)
Utilizando o resultado do item anterior fica mais fácil. Poderemos então utilizar a capacidade dos canais binários simétricos $C_1 = 1 - H(\epsilon)$ e $C_2 = 1 - H(\delta)$. Desta forma, definimos uma função $f(\alpha) := H(\alpha) + \alpha C_1 + (1 - \alpha)C_2$, a qual queremos maximizar com relação ao parâmetro α. Teremos então um problema de otimização unidimensional. Devemos encontrar o ponto em que a derivada primeira se igual a zero, quando a derivada segunda é negativa.

...

Canal da Soma XI

Exercício (Canal da Soma)

continuação...

$$\begin{aligned}\frac{df(\alpha)}{d\alpha} &= H'(\alpha) + C_1 - C_2 \\ &= \frac{1}{\ln 2}\ln\left(\frac{1-\alpha}{\alpha}\right) + \frac{1}{\ln 2}C_1 - \frac{1}{\ln 2}C_2 \\ &= \log\left(\frac{1-\alpha}{\alpha}\right) + C_1 - C_2 \qquad (714)\end{aligned}$$

...

Canal da Soma XII

Exercício (Canal da Soma)

continuação...

$$\begin{aligned}\frac{d^2 f(\alpha)}{d\alpha^2} &= \frac{1}{\ln 2}\frac{\alpha}{1-\alpha}\frac{-\alpha-1+\alpha}{\alpha^2} \\ &= \frac{1}{\ln 2}\frac{1}{\alpha-1}\end{aligned} \tag{715}$$

como $0 \leq \alpha \leq 1$, *teremos que* $\frac{d^2 f(\alpha)}{d\alpha^2} < 0$, *logo o ponto que encontraremos quando* $\frac{df(\alpha)}{d\alpha} = 0$ *será um ponto de máximo.*

...

Canal da Soma XIII

Exercício (Canal da Soma)

Canal da Soma XIV

O ponto de máximo será então dado por

$$
\begin{aligned}
\log\left(\frac{1-\alpha}{\alpha}\right) + C_1 - C_2 &= 0 \\
\log\left(\frac{1-\alpha}{\alpha}\right) &= C_2 - C_1 \\
\frac{1-\alpha}{\alpha} &= \frac{2^{C_2}}{2^{C_1}} \\
\alpha 2^{C_2} &= (1-\alpha)2^{C_1} \\
\alpha(2^{C_1} + 2^{C_2}) &= 2^{C_1} \\
\alpha &= \frac{2^{C_1}}{2^{C_1} + 2^{C_2}} \qquad (716)
\end{aligned}
$$

...

Canal da Soma XV

Exercício (Canal da Soma)

continuação...

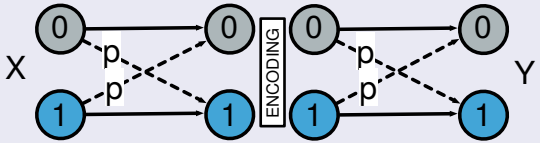
Substituindo o valor encontrado para α, teremos

$$\begin{aligned} C &= H(\alpha) + \alpha C_1 + (1-\alpha)C_2 \\ &= -\alpha \log \alpha - (1-\alpha)\log(1-\alpha) + \alpha C_1 + (1-\alpha)C_2 \\ &= \alpha(C_1 - C_1 + \log(2^{C_1} + 2^{C_2}) + (1-\alpha)(C_2 - C_2 + \log(2^{C_1} - 2^{C_2})) \\ &= \alpha \log(2^{C_1} + 2^{C_2}) + (1-\alpha)\log(2^{C_1} + 2^{C_2}) \\ &= \log(2^{C_1} + 2^{C_2}) \end{aligned} \tag{717}$$

Canais em Cascata com Codificador I

Exercício (Canais em cascata com codificador entre eles)

Considere a canais binários simétricos conectados em cascata com um codificador entre eles, conforme ilustrado abaixo. Calcule a capacidade do canal entre X e Y.



...

Canais em Cascata com Codificador II

Exercício (Canais em cascata com codificador entre eles)

continuação...

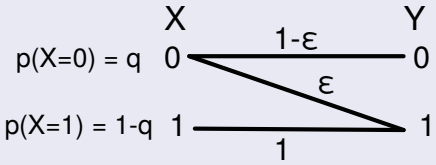
(solução)

Os símbolos são re-codificados após o primeiro canal, desta forma a capacidade torna-se $C = \min(C_1, C_2)$*, onde* C_1 *e* C_2 *são as capacidades do primeiro e segundo canais binários, respectivamente. Como neste caso os canais são iguais, temos* $C_1 = C_2 = 1 - H(p)$*. Logo, a capacidade do canal será* $C = 1 - H(p)$*, sendo este valor alcançado quando a distribuição de* X *for uniforme.*

Canal Z I

Exercício (Canal Z)

Considere o canal Z, com capacidade C_Z, ilustrado abaixo. Considere a seguinte distribuição de entrada: $p(X = 0) = q$, $p(X = 1) = 1 - q$. Encontre a equação que deverá ser otimizada para encontrarmos a capacidade deste canal (simplifique a equação ao máximo, mas não é necessário solucioná-la).



obs.: as características de erro em sistemas ópticos e memória de alguns semi-condutores podem ser modeladas através de um canal z.

...

Canal Z II

Exercício (Canal Z)

continuação...

(solução)

Devemos encontrar q que maximiza $I(X;Y) = H(Y) - H(Y|X)$. Primeiramente iremos encontrar uma expressão para $I(X;Y)$ em termos de q e ϵ. Vamos tomar a derivada com relação a q e achar o q que faz com que esta derivada seja igual a zero.

Note que $p(Y = 0) = q(1 - \epsilon)$ e, por conseguinte, $p(Y = 1) = 1 - q(1 - \epsilon)$, assim, $H(Y) = H(q(1 - \epsilon))$. ...

Canal Z III

Exercício (Canal Z)

continuação...

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(q(1-\epsilon)) + (1-q)\underbrace{H(Y|X=1)}_{=0} + q\underbrace{H(Y|X=0)}_{=H(\epsilon)} \\ &= H(q(1-\epsilon)) + qH(\epsilon) \\ &= -q(1-\epsilon)\log q(1-\epsilon) - (1-q(1-\epsilon))\log(1-q(1-\epsilon)) + \\ &\quad qH(\epsilon) \end{aligned} \tag{718}$$

Devemos tomar $\frac{dI(X;Y)}{dq} = 0$ *e achar* q *que satisfaz esta equação.*

...

Canal Z IV

Exercício (Canal Z)

continuação...

$$C \approx 1 - \frac{1}{2} H(\epsilon) \tag{719}$$

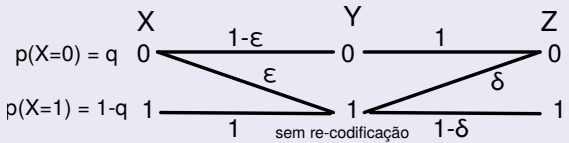
que será alcançada quando

$$q = \frac{1}{(1-\epsilon)(1+2^{H(\epsilon)/(1-\epsilon)})} \tag{720}$$

Canal Z V

Exercício (Canal Z em cascata)

Considere agora a cascata de dois canais Z, conforme ilustrado na figura abaixo.



A saída de um canal é inserida diretamente no canal seguinte, sem recodificação. Encontre as probabilidades de transição do canal final criado pela cascata dos 2 canais Z.

...

Canal Z VI

Exercício (Canal Z em cascata)

continuação...

a) *Encontre as probabilidades de transição para o canal final (entre X e Z).*

(solução)

O canal efetivo entre X e Z é descrito pelas probabilidades

$$p(Z=0|X=0) = (1-\epsilon) + \epsilon\delta \tag{721}$$

$$p(Z=1|X=0) = \epsilon(1-\delta) \tag{722}$$

$$p(Z=0|X=1) = \delta \tag{723}$$

$$p(Z=1|X=1) = 1-\delta \tag{724}$$

...

Canal Z VII

Exercício (Canal Z em cascata)

continuação...

$$p(z|x) = \begin{pmatrix} (1-\epsilon)+\epsilon\delta & \epsilon(1-\delta) \\ \delta & 1-\delta \end{pmatrix} \tag{725}$$

...

Canal Z VIII

Exercício (Canal Z em cascata)

continuação...

b) *Encontre o valor de δ (em termos de ϵ) que faz com que o canal final XZ seja simétrico e encontre a capacidade deste canal C_{XZ}.*

(solução)
Devemos ter $p(Z=1|X=0) = p(Z=0|X=1)$, ou seja,

$$\epsilon = \frac{\delta}{1-\delta} \tag{726}$$

Note que esta escolha levará também a $p(Z=1|X=1) = p(Z=0|X=0)$.

...

Canal Z IX

Exercício (Canal Z em cascata)

continuação...

Como o canal é simétrico, teremos

$$\begin{aligned} C &= \log|\mathcal{Z}| - H(r) \\ &= \log 2 - H(\delta) \\ &= 1 - H(\delta). \end{aligned} \tag{727}$$

onde r é uma linha da matriz de transição, ou seja, $r = [\delta \quad (1-\delta)]$, utilizando $\epsilon = \delta/(1-\delta)$.

...

Canal Z X

Exercício (Canal Z em cascata)

continuação...

c) *Considere agora que, em Y, seja possível decodificar e recodificar a sequência recebida. Qual é a capacidade do sistema agora?*

(solução)

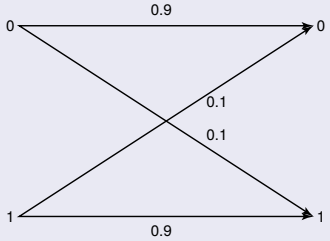
$$C = \min(C_Z, C_{YZ}). \tag{728}$$

onde C_Z é a capacidade do canal X—Y e C_{YZ} é a capacidade do canal Y—Z.

Sequencias Típicas I

Exercício (Sequencias Típicas)

Vamos calcular o conjunto de pares de variáveis aleatórias conjuntamente típicas e conectadas através de um canal binário simétrico e a probabilidade de erro para a decodificação através da tipicidade conjunta para este canal.



...

Sequencias Típicas II

Exercício (Sequencias Típicas)

continuação...

Vamos considerar um canal binário simétrico com probabilidade de troca de bit de 0.1. *A distribuição de entrada que atinge o limite da capacidade é a distribuição uniforme, i.e.,* $p(x) = (1/2, 1/2)$, *que leva à seguinte distribuição conjunta* $p(x, y)$ *para este canal:*

$X \setminus Y$	0	1
0	0.45	0.05
1	0.05	0.45

A distribuição marginal de Y *também será* $(1/2, 1/2)$.

...

Sequencias Típicas III

Exercício (Sequencias Típicas)

continuação...

a) *Calcule* $H(X)$, $H(Y)$, $H(X,Y)$ *e* $I(X;Y)$ *para a distribuição dada.*

(solução)
$H(X) = H(Y) = 1$ *bit já que ambos possuem distribuição* $(1/2, 1/2)$.
$H(X,Y) = H(X) + H(Y|X) = 1 + H(p) = 1 - 0.9\log 0.9 - 0.1\log 0.1 = 1 + 0.469 = 1.469$ *bits*. $I(X;Y) = H(Y) - H(Y|X) = 1 - 0.469 = 0.531$ *bits*.

...

Sequencias Típicas IV

Exercício (Sequencias Típicas)

continuação...

b) *Sejam $X_1, X_2, \ldots, X_n$, i.i.d. com distribuição de Bernoulli$(1/2)$. Dentre as 2^n possíveis sequências de entrada de comprimento n, quais delas são típicas, i.e., membros de $A_\epsilon^{(n)}(X)$ para $\epsilon = 0.2$? Quais são as sequências típicas em $A_\epsilon^{(n)}(Y)$?*

(solução)

No caso em que a distribuição é uniforme, toda sequência terá a mesma probabilidade $(1/2)^n$ e, desta forma, para toda sequencia, $-\frac{1}{n}\log p(x^n) = -\frac{1}{n} n \log 1/2 = 1 = H(X)$. Desta forma teremos que toda sequência é típica, i.e. $\in A_\epsilon^{(n)}$.

De forma semelhante, toda sequência y^n é típica, i.e., $\in A_\epsilon^{(n)}(Y)$.

...

Sequencias Típicas V

Exercício (Sequencias Típicas)

continuação...

c) *O conjunto das sequências conjuntamente típicas $A_\epsilon^{(n)}(X,Y)$ é definido pelo conjunto das sequências que satisfazem a Equação 612. As duas primeiras equações correspondem à condição que x^n e y^n estejam em $A_\epsilon^{(n)}(X)$ e $A_\epsilon^{(n)}(Y)$ respectivamente. Considere a última condição, que pode ser reescrita da seguinte forma: $-\frac{1}{n}\log p(x^n, y^n) \in (H(X,Y)-\epsilon, H(X,Y)+\epsilon)$. Seja k o número de lugares em que a sequência x^n difere de y^n (k é uma função de ambas sequências).*

...

Sequencias Típicas VI

Exercício (Sequencias Típicas)

continuação...

Podemos escrever então:

$$\begin{aligned} p(x^n, y^n) &= \prod_{i=1}^{n} p(x_i, y_i) \\ &= (0.45)^{n-k}(0.05)^k \\ &= \left(\frac{1}{2}\right)^n (1-p)^{n-k} p^k \end{aligned} \tag{729}$$

...

Sequencias Típicas VII

Exercício (Sequencias Típicas)

continuação...

Uma forma alternativa de analisar esta probabilidade é ver o canal binário simétrico como um canal aditivo $Y = X \oplus Z$, onde Z é uma v.a. binária igual a 1 com probabilidade p e independente de X. Neste caso,

$$\begin{aligned} p(x^n, y^n) &= p(x^n)p(y^n|x^n) \\ &= p(x^n)p(z^n|x^n) \\ &= p(x^n)p(z^n) \\ &= \left(\frac{1}{2}\right)^n (1-p)^{n-k}p^k \end{aligned} \tag{730}$$

Mostre que a condição que faz (x^n, y^n) ser conjuntamente típico é equivalente à condição de que x^n seja típico e $z^n = y^n - x^n$ seja típico.

...

Sequencias Típicas VIII

Exercício (Sequencias Típicas)

Sequencias Típicas IX

continuação...

(solução)

As condições para $(x^n, y^n) \in A_\epsilon^{(n)}(X,Y)$ *são*

$$
\begin{aligned}
A_\epsilon^{(n)} \quad = \quad & \{(x_{1:n}, y_{1:n}) \in \mathcal{X}^n \times \mathcal{Y}^n : \\
& \textit{a)} \quad |-\frac{1}{n}\log p(x^n) - H(X)| < \epsilon, \quad \textit{típico em x} \\
& \textit{b)} \quad |-\frac{1}{n}\log p(y^n) - H(Y)| < \epsilon, \quad \textit{típico em y} \\
& \textit{c)} \quad |-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)| < \epsilon, \quad \textit{típico em (x,y)} \\
& \} \qquad (731)
\end{aligned}
$$

...

Sequencias Típicas X

Exercício (Sequencias Típicas)

continuação...

Mas, como dito, cada sequência x^n e y^n satisfaz as duas primeiras condições. Desta forma, a única condição que importa é a última. Como dito anteriormente,

$$\begin{aligned} -\frac{1}{n}\log p(x^n, y^n) &= -\frac{1}{n}\log\left(\left(\frac{1}{2}\right)^n p^k (1-p)^{n-k}\right) \\ &= 1 - \frac{k}{n}\log p - \frac{n-k}{n}\log(1-p) \end{aligned} \tag{732}$$

...

Sequencias Típicas XI

Exercício (Sequencias Típicas)

continuação...

Então o par (x^n, y^n) *é conjuntamente típico se e somente se* $|1 - \frac{k}{n} \log p - \frac{n-k}{n} \log(1-p) - H(X,Y)| < \epsilon$, *isto é, se e somente se* $| - \frac{k}{n} \log p - \frac{n-k}{n} \log(1-p) - H(p)| < \epsilon$, *ou seja,* p *é próximo de* $\frac{k}{n}$, *e também é exatamente a condição para que* $z^n = y^n \oplus x^n$ *seja típico. Então o conjunto de pares* (x^n, y^n) *conjuntamente típicos é o conjunto tal que o número de lugares em que as sequências* x^n *e* y^n *diferem é próximo de* np.

...

Sequencias Típicas XII

Exercício (Sequencias Típicas)

continuação...

d) *Podemos agora calcular o tamanho do conjunto* $A_\epsilon^{(n)}(Z)$ *para* $n = 25$ *e* $\epsilon = 0.2$. *Abaixo segue uma tabela com as probabilidades e número de sequências com* k *uns.*

...

Sequencias Típicas XIII

Exercício (Sequencias Típicas)

Sequencias Típicas XIV

continuação...

k	$\binom{n}{k}$	$\sum_{j \leq k} \binom{n}{j}$	$p(x^n) = p^k(1-p)^{n-k}$	$\binom{n}{k} p^k(1-p)^{n-k}$	Cumul. pr.	$-\frac{1}{n} \log p(x^n)$
0	1	1	7.178975e-02	0.071790	0.071790	0.152003
1	25	26	7.976639e-03	0.199416	0.271206	0.278800
2	300	326	8.862934e-04	0.265888	0.537094	0.405597
3	2300	2626	9.847704e-05	0.226497	0.763591	0.532394
4	12650	15276	1.094189e-05	0.138415	0.902006	0.659191
5	53130	68406	1.215766e-06	0.064594	0.966600	0.785988
6	177100	245506	1.350851e-07	0.023924	0.990523	0.912785
7	480700	726206	1.500946e-08	0.007215	0.997738	1.039582
8	1081575	1807781	1.667718e-09	0.001804	0.999542	1.166379
9	2042975	3850756	1.853020e-10	0.000379	0.999920	1.293176
10	3268760	7119516	2.058911e-11	0.000067	0.999988	1.419973
11	4457400	11576916	2.287679e-12	0.000010	0.999998	1.546770
12	5200300	16777216	2.541865e-13	0.000001	0.999999	1.673567

Figura 21: As sequências com mais de 12 uns foram omitidas pois a probabilidade delas é desprezível (e não estão no conjunto típico).

Codificador e Decodificador como parte do Canal I

Exercício (Codificador e Decodificador como parte do Canal)

Considere um canal binário simétrico com probabilidade de crossover 0.1. Um possível esquema de codificação para este canal com duas palavras de comprimento 3 é codificar a mensagem a_1 como 000 e a_2 como 111. Com este esquema de codificação, podemos considerar a combinação do codificador, canal e decodificador como formando um novo canal binário simétrico com duas entradas a_1 e a_2 e duas saídas a_1 e a_2.

...

Codificador e Decodificador como parte do Canal II

Exercício (Codificador e Decodificador como parte do Canal)

continuação...

a) *Calcule a probabilidade de crossover deste canal.*

(solução)

Neste novo canal não haverá crossover quando nenhum dos 3 bits forem trocados ou quando apenas um dos 3 bits forem trocados. Iremos calcular então

$$\Pr(\textit{crossover}) = 1 - \Pr(\textit{não crossover}) = 1 - ((1-p)(1-p)(1-p) + 3p(1-p)(1-p)) = 0.028.$$

Desta forma, a probabilidade de crossover neste novo canal será 0.028.

...

Codificador e Decodificador como parte do Canal III

Exercício (Codificador e Decodificador como parte do Canal)

continuação...

b) *Qual é a capacidade deste canal em bits por transmissão do canal original?*

(solução)

A capacidade de um canal binário simétrico com probabilidade de crossover p é dada por $1 - H(p)$. Neste caso, temos $p = 0.028$ e assim $1 - H(0.028) = 1 - 0.18426 = 0.81574$ bits para cada palavra de 3 bits. Isto corresponde a $0.81574/3 = 0.27191$ bits por transmissão do canal original.

...

Codificador e Decodificador como parte do Canal IV

Exercício (Codificador e Decodificador como parte do Canal)

continuação...

c) *Qual é a capacidade do canal binário simétrico original com probabilidade de crossover 0.1?*

(solução)

A capacidade do canal original seria de $1 - H(0.1) = 0.531$ *bits/transmissão.*

...

Codificador e Decodificador como parte do Canal V

Exercício (Codificador e Decodificador como parte do Canal)

continuação...

d) *Prove um resultado geral mostrando que para qualquer canal, considerando o codificador e decodificador conjuntamente como um novo canal de mensagens para mensagens estimadas não irá aumentar a capacidade em bits por transmissão do canal original.*

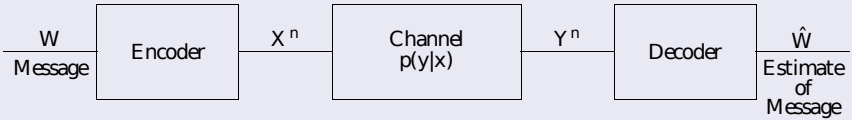
...

Codificador e Decodificador como parte do Canal VI

Exercício (Codificador e Decodificador como parte do Canal)

continuação...

(solução)



Pela desigualdade de processamento de dados temos $I(W;\hat{W}) \leq I(X^n;Y^n)$ *e assim*

$$C_W = \max_{p(w)} I(W;\hat{W}) \leq \frac{1}{n} \max_{p(x^n)} I(X^n;Y^n) = C. \tag{734}$$

Então a capacidade do canal por transmissão não aumentará ao considerar o codificador e decodificador como parte do canal.

Códigos e Codificação I

- O teorema de Shannon diz que existe uma sequência de códigos tais que: se $R < C$, então a probabilidade de erro vai a zero.
- Ele não fornece um código ou uma maneira de encontrá-lo.
- Codificação típica não é prática, pois haverá formação de blocos com tamanho exponencialmente grandes.
- Devemos adicionar redundância suficiente à mensagem para que a mensagem original seja decodificada de forma não ambígua.

Soluções Físicas I

- Podemos estabelecer uma comunicação mais confiável alterando as características físicas do meio, reduzindo assim o ruído interferente (diminuir p em um canal binário simétrico).
- Utilizar circuitos implementados com componentes de menor tolerância.
- Melhorar as condições do ambiente (condições térmicas, poeira e ar).
- Utilizar mais área/volume físico por bit.
- Utilizar maior potência na transmissão, fazendo que o ruído seja menos significativo.

Códigos de Repetição I

- cada símbolo é repetido k vezes
- mensagem: $x_1, x_2, \ldots, x_n$
- transmitido: $\underbrace{x_1 x_1 \ldots x_1}_{k \text{ vezes}} \underbrace{x_2 x_2 \ldots x_2}_{k \text{ vezes}} \ldots \underbrace{x_n x_n \ldots x_n}_{k \text{ vezes}}$.
- Para muitos canais (por exemplo: canal binário simétrico com $p < 1/2$), o erro tende a zero quando $k \to \infty$.
- Decodificação simples: quando k é ímpar, utilizar o voto da maioria (que é ótimo para canal binário simétrico).
- $R \propto 1/k \to 0$ quando $k \to \infty$.
- Veja o exemplo apresentado previamente 125.
- Estamos supondo que o ruído seja branco, mas algumas vezes o ruído pode ter outra característica, como por exemplo um ruído em rajadas. Neste caso, o código de repetição seria desastroso. Poderíamos adaptá-lo intercalando os símbolos de forma a minimizar o efeito nocivo de uma rajada.

Códigos de Repetição II

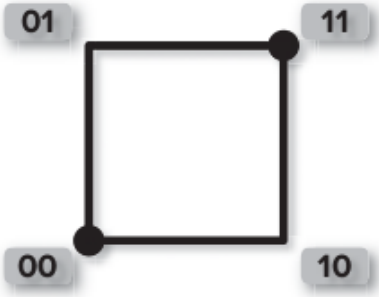


Figura 22: Código de repetição com $k = 2$ ($d_H = 1$).

Códigos de Repetição III

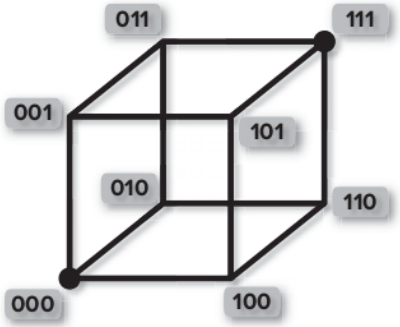


Figura 23: Código de repetição com $k = 3$ ($d_H = 3$).

Se d_H (distância de Hamming) for par, podemos detectar $d_H/2$ erros e corrigir $d_H/2 - 1$ erros.
Se d_H for ímpar, podemos corrigir até $(d_H - 1)/2$ erros.

Código de Verificação de Paridade Simples I

- Entrada/saída binária: $\mathcal{X} = \mathcal{Y} = \{0, 1\}$.
- Blocos de comprimento $n-1$ bits: $x_{1:n-1}$.
- O n-ésimo bit é um indicador do número ímpar de bits iguais a 1 em $x_{1:n-1}$, ou seja,

$$x_n \leftarrow \mod\left(\sum_{i=1}^{n-1} x_i, 2\right). \tag{735}$$

- Uma condição necessária para que uma palavra seja válida é

$$\mod\left(\sum_{i=1}^{n} x_i, 2\right) = 0. \tag{736}$$

- Se ocorrer um número ímpar de erros, esta condição não será satisfeita. Podemos detectar que ocorreu um número ímpar de erros.
- Um número par de erros não será detectado.

Código de Verificação de Paridade Simples II

- Só conseguimos detectar alguns erros e não conseguimos corrigir erros.
- Verificação de paridade é a base de vários esquemas mais sofisticados de codificação (por exemplo: verificação de paridade de baixa densidade (*low-density parity check*, LDPC), código de Hamming, etc.).

Código de Hamming $(7, 4, 3)$ I

- comprimento da palavra: 7; número de bits de informação: 4; número de bits de verificação de paridade 3.
- $\mathcal{X} = \mathcal{Y} = \{0, 1\}$.
- $R = 4/7$ bits por utilização do canal.
- Bits de dados: $x_0, x_1, x_2, x_3 \in \{0, 1\}$.
- Bits de redundância: x_4, x_5, x_6.
- Os bits de paridade são determinados pelas equações:

$$x_4 = (x_1 + x_2 + x_3) \mod 2 \tag{737}$$

$$x_5 = (x_0 + x_2 + x_3) \mod 2 \tag{738}$$

$$x_6 = (x_0 + x_1 + x_3) \mod 2 \tag{739}$$

- Exemplo: $(x_0, x_1, x_2, x_3) = (0110)$, então $(x_4, x_5, x_6) = (011)$ e assim a palavra de 7 bits será (0110011).

Código de Hamming $(7, 4, 3)$ II

- Isto pode ser visto na forma matricial como:

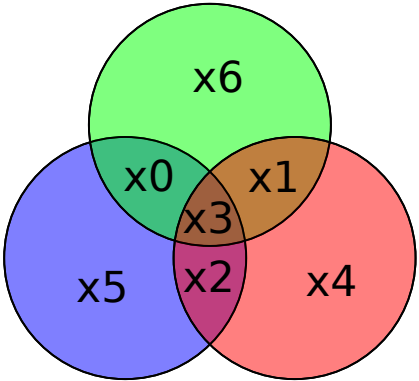
$$\mathbf{G}p = x \tag{740}$$

onde p é o vetor de dados, $p = (x_0, x_1, x_2, x_3)$, e $\mathbf{G}$ é a matriz geradora do código de Hamming e x os dados com os bits de paridade.

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}. \tag{741}$$

- Representação gráfica

Código de Hamming $(7, 4, 3)$ III



Código de Hamming $(7, 4, 3)$ IV

- Podemos também descrever os bits de paridade através das seguintes equações

$$\begin{array}{cccccccc} & x_1 & +x_2 & +x_3 & +x_4 & & & = 0 \\ x_0 & & +x_2 & +x_3 & & +x_5 & & = 0 \\ x_0 & +x_1 & & +x_3 & & & +x_6 & = 0 \end{array} \tag{742}$$

- Alternativamente, podemos escrever $\mathbf{H}x = 0$, onde $x^T = (x_0, x_1, \ldots, x_6)$ e

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{743}$$

- As palavras estão no espaço nulo (núcleo)[3] de $\mathbf{H}$.

Código de Hamming $(7, 4, 3)$ V

- Note que as colunas de $\mathbf{H}$ são todas as 7 possíveis permutações de colunas de tamanho 3 não-nulas.
- As palavras são definidas pelo espaço nulo de $\mathbf{H}$, i.e. $\{x : \mathbf{H}x = 0, x \neq 0\}$.
- Como o posto[4] da $\mathbf{H}$ é 3, o espaço nulo será de tamanho 4, visto que $\mathbf{H}$ possui 7 colunas [5]. Esperamos assim encontrar $2^4 = 16$ vetores binários neste espaço nulo.

[3] O núcleo (espaço nulo) de uma transformação linear $L : V \to W$ entre dois espaços vetoriais é o conjunto de todos os elementos de $v \in V$ para os quais $L(v) = 0$, isto é

$$\ker(L) = \{v \in V : L(v) = 0\}, \tag{744}$$

onde 0 é o vetor nulo em W.

[4] O posto (*rank*) de uma matriz é a dimensão do espaço vetorial gerado pelas colunas da matriz. Isto é o mesmo que a dimensão do espaço gerado pelas linhas. O posto é uma medida de 'não-degeneração' do sistemas de equações lineares e transformação linear codificada pela matrix.

[5] O teorema do posto-nulidade diz que o posto e a nulidade de uma matriz somados devem ser igual ao número de colunas da matriz. Se $\mathbf{A}$ é uma matriz $m \times n$, então devemos ter

$$\text{rank}(\mathbf{A}) + \text{null}(\mathbf{A}) = n \tag{745}$$

Código de Hamming $(7, 4, 3)$ I

- 16 vetores no espaço nulo de $\mathbf{H}$:

$$\begin{array}{cccc} 0000000 & 0100101 & 1000011 & 1100110 \\ 0001111 & 0101010 & 1001100 & 1101001 \\ 0010110 & 0110011 & 1010101 & 1110000 \\ 0011001 & 0111100 & 1011010 & 1111111 \end{array} \tag{746}$$

- Os 4 primeiros bits são os valores variando de 0 a 15 em binário (todas as *strings* binárias de comprimento 4), são os bits de dados. Os 3 bits em sequência são os bits de redundância (paridade).
- Uma palavra válida do código deve ser um dos vetores acima, $C = \{x : \mathbf{H}x = 0\}$.
- Se $v_1, v_2 \in C$ então $\mathbf{H}(v_1 + v_2) = \mathbf{H}v_1 + \mathbf{H}v_2 = 0$, desta forma, $v_1 + v_2 \in C$.
- Da mesma forma, $v_1 - v_2 \in C$.
- O conjunto de palavras (*codewords*) é um conjunto fechado sob a adição e subtração.
- O número mínimo de 1s nestas palavras é 3. Isto é chamado de peso do código.

Código de Hamming $(7, 4, 3)$ II

- Por que 3? Suponha que o peso do código seja 2 com elementos não nulos nas posições i e j. Neste caso, por ser uma palavra do código, devemos ter $\mathbf{H}x = 0$, assim a soma da i-ésima e j-ésima colunas de $\mathbf{H}$ deverá ser igual a zero. Como as colunas de $\mathbf{H}$ são todas diferentes, a soma de quaisquer duas colunas é não nula. Desta forma não é possível que o código tenha peso 2.
- Não podemos ter peso 1 pois as palavras com peso 1 não estão no espaço nulo, pois não existe uma coluna nula em $\mathbf{H}$.
- Peso 3 é possível, pois a soma de duas colunas é igual a uma outra coluna e a soma de dois vetores binários iguais é igual a zero ($\mod 2$).

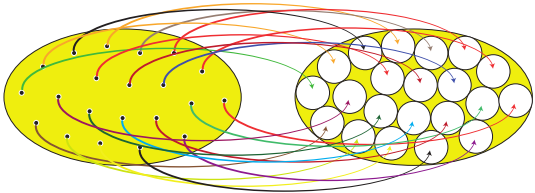
- A distância mínima entre palavras deste código também é 3, que é o número mínimo de diferenças entre as palavras. Se $v_1, v_2 \in C = \{x : \mathbf{H}x = 0\}$, então $d_H(v_1, v_2) \geq 3$ onde

$$d_H(x, y) = \sum_i \mathbf{1}_{\{x(i) \neq y(i)\}} \tag{747}$$

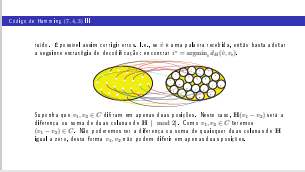
é a distância de Hamming. Note que, quanto maior a distância entre as palavras de um código, menor a chance de haver confusão se a mensagem enviada for corrompida por

Código de Hamming $(7, 4, 3)$ III

ruído. É possível assim corrigir erros. I.e., se $\hat{v}$ é uma palavra recebida, então basta adotar a seguinte estratégia de decodificação: encontrar $i^* = \text{argmin}_i \, d_H(\hat{v}, v_i)$.

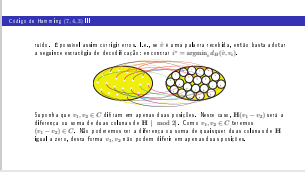


Suponha que $v_1, v_2 \in C$ difiram em apenas duas posições. Neste caso, $\mathbf{H}(v_1 - v_2)$ será a diferença ou soma de duas colunas de $\mathbf{H}$ ($\mod 2$). Como $v_1, v_2 \in C$ teremos $(v_1 - v_2) \in C$. Não poderemos ter a diferença ou soma de quaisquer duas colunas de $\mathbf{H}$ igual a zero, desta forma v_1, v_2 não podem diferir em apenas duas posições.



A distância de Hamming entre duas palavras pode ser calculada fazendo um XOR entre elas e contabilizando o número de uns.

As propriedades de um código de bloco de detectar e corrigir erros dependem da distância de Hamming mínima entre palavras deste código. Para detectar d erros de forma confiável, precisamos de uma distância mínima de $d + 1$ entre as palavras do código, número de trocas de bits necessária para gerar uma outra palavra válida no código. Da mesma forma, para corrigir d erros, precisamos de uma distância mínima de $2d + 1$. Assumimos o pressuposto de que um maior número de erros é menos provável e assim corrige-se o erro escolhendo a palavra válida mais próxima (menor número de bits trocado).



Seja um código com mensagens de comprimento $n = m + r$, com m bits de dados e r bits de paridade. Vamos analisar o caso de corrição de um único bit. Cada palavra possui n vizinhos inválidos e assim $n + 1$ padrões associados a ela (n inválidos e a palavra do código). São 2^m palavras no código. O total de sequências possíveis com n bits é 2^n. Demos ter então $(n+1)2^m \leq 2^n$. Usando $n = m+r$, teremos

$$(m + r + 1) \leq 2^r. \tag{748}$$

Dado m podemos calcular o número de bits de verificação necessários para corrigir um único erro.

Código de Hamming - Canal Binário Simétrico I

- A decodificação de máxima verosimilhança requer o conhecimento das características do canal e ainda é um problema NP-difícil.
- Vamos assumir que temos um BSC(p) (canal binário simétrico com probabilidade de troca p).
- $x = (x_0, x_1, \ldots, x_6)$ é transmitido, e será recebido

$$y = x + z = (x_0 + z_0, x_1 + z_1, \ldots, x_6 + z_6), \tag{749}$$

  onde $z = (z_0, z_1, \ldots, z_6)$ é o vetor de ruido aditivo.
- Recebemos y e queremos determinar x. Iremos calcular a síndrome de y

$$s = \mathbf{H}y = \mathbf{H}(x + z) = \underbrace{\mathbf{H}x}_{=0} + \mathbf{H}z = \mathbf{H}z \tag{750}$$

- Se $s = 0$, então todos as verificações de paridade são satisfeitas por y, o que é uma condição necessária para que tenhamos uma palavra correta.

Código de Hamming - Canal Binário Simétrico II

- $s = \mathbf{H}z$ é uma combinação linear das colunas de $\mathbf{H}$

$$s = z_0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + z_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + z_2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \ldots + z_6 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{751}$$

- Como $y = x + z$, basta encontrar z para determinar x, já que y é conhecido.
- Precisamos resolver $s = \mathbf{H}z$, um sistema de 3 equações e 7 variáveis. Teremos 4 graus de liberdade. Como as variáveis são binárias, teremos $2^4 = 16$ possíveis soluções.
- Exemplo: Suponha que $y^T = 0111001$ seja recebido (não é uma palavra válida), então calcularemos a síndrome de y, $s = \mathbf{H}y = \mathbf{H}z = (101)^T$ e as 16 possíveis soluções para z são

$$\begin{array}{cccc} 0100000 & 0010011 & 0101111 & 1001001 \\ 1100011 & 0001010 & 1000110 & 1111010 \\ 0000101 & 0111001 & 1110101 & 0011100 \\ 0110110 & 1010000 & 1101100 & 1011111 \end{array} \tag{752}$$

Código de Hamming - Canal Binário Simétrico III

- Dentre os 128 possíveis vetores que poderíamos ter, restringimos a apenas 16.
- Qual é a probabilidade de cada possível solução? Assumindo que temos um canal binário simétrico com $p < 1/2$, a solução mais provável é aquela com menor peso. Qualquer solução com peso k possui probabilidade p^k.
- Note que existe apenas uma possível solução com peso 1. Esta é a solução mais provável.
- A solução mais provável, para o exemplo, é $z = (0100000)^T$ e assim teremos $y = x + z$, como $y = (0111001)^T$ teremos a palavra $x = (0011001)^T$, assim os bits de informação são 0011.
- Para qualquer s, existe uma única solução de peso mínimo para z em $s = \mathbf{H}z$.
- Se $s = (000)^T$, então a única solução é $z = (0000000)^T$.
- Para uma solução de peso 1, qualquer outro s será igual a uma das colunas de $\mathbf{H}$, poderemos assim gerar z fazendo o bit correspondente igual a 1.

Procedimento de Decodificação de Hamming I

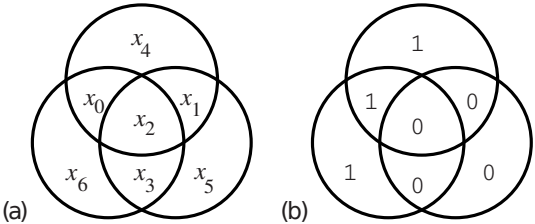
Procedimento de decodificação de síndrome para um dado y recebido:

1) Calcular a síndrome $s = \mathbf{H}y$.
2) Se $s = (000)^T$, faça $z \leftarrow (0000000)^T$ e vá ao passo 4.
3) Caso contrário, localize a única coluna de $\mathbf{H}$ igual a s, faça z um vetor cheio de zeros mas com 1 na posição correspondente.
4) Faça $x \leftarrow y + z$
5) saída: (x_0, x_1, x_2, x_3)

Este procedimento é capaz de corrigir um único bit de erro, mas falha quando há mais do que um único bit de erro.

Visualização do procedimento de decodificação I

- O procedimento de decodificação pode ser visualizado através de um diagrama de Venn.



- Os bits de dados (x_0, x_1, x_2, x_3) estão nas intersecções e fora delas os bits de paridade (x_4, x_5, x_6).

Visualização do procedimento de decodificação II

- Dentro de cada círculo devemos ter um número par de bits iguais a 1, indicando que não há erro.

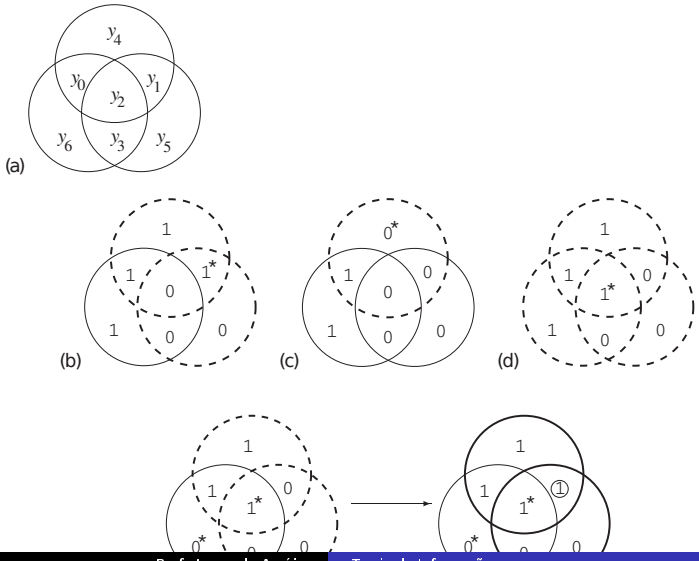
$$x_4 \equiv (x_0 + x_1 + x_2) \mod 2 \tag{753}$$

$$x_5 \equiv (x_1 + x_2 + x_3) \mod 2 \tag{754}$$

$$x_6 \equiv (x_0 + x_2 + x_3) \mod 2 \tag{755}$$

- A síndrome pode ser vista como o caso em que a condição de paridade não é satisfeita.
- Foi dito que, para $s \neq (0, 0, 0)$ sempre existe um *flip* de bit que irá fazer com que todas condições de paridade sejam satisfeitas.

Visualização do procedimento de decodificação III



Codificação I

- Existem outros algoritmos de codificação.
- Códigos de Reed Salomon
- Códigos de Bose, Ray-Chaudhuri, Hocquenghem
- Códigos Convolucionais
- Códigos Turbo
- Códigos de Verificação de Baixa Densidade
- Todos desenvolvidos na busca de obtermos bons códigos com baixa taxa e atingir o limite de Shannon.

Código Hamming I

Exercício (Código Hamming)

Utilizando o código de Hamming (7,4,3), realize a decodificação das sequências recebidas abaixo utilizando para tanto o procedimento de decodificação de Hamming.

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{756}$$

...

Código Hamming II

Exercício (Código Hamming)

continuação...

a) $y = 1101011$

(solução)
calculando a síndrome teremos

$$s = \mathbf{H}y = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \tag{757}$$

...

Código Hamming III

Exercício (Código Hamming)

continuação...

$s = \mathbf{H}z$ é uma combinação linear das colunas de $\mathbf{H}$

$$s = z_0 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + z_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + z_2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + \ldots + z_6 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{758}$$

as 16 possíveis soluções seriam:

$$\begin{array}{cccc} 0101000 & 1011000 & 1100100 & 0010100 \\ 0000010 & 1110010 & 1001110 & 0111110 \\ 1000001 & 0110001 & 0001101 & 1111101 \\ 1101011 & 0011011 & 0100111 & 1010111 \end{array} \tag{759}$$

...

Código Hamming IV

Exercício (Código Hamming)

continuação...

vamos escolher aquela com peso 1, basta fazer z um vetor nulo e inserir 1 apenas na posição correspondente à coluna de $\mathbf{H}$ igual a s

$$z = (0000010)^T \tag{760}$$

Assim, iremos decodificar

$$\hat{x} = y + z = 1101011 + 0000010 = 1101001 \tag{761}$$

podemos verificar que $\mathbf{H}x = 0$

...

Código Hamming V

Exercício (Código Hamming)

continuação...

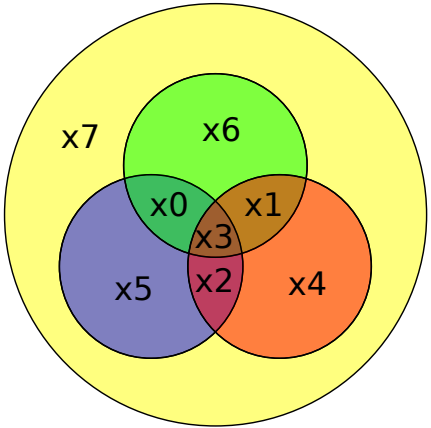
b) $y = 0110110$

(solução)
síndrome $s = (101)^T$
$z = 0100000$

$$\hat{x} = y + z = 0110110 + 0100000 = 0010110 \tag{762}$$

Código de Hamming Estendido I

O Código de Hamming $(7, 4, 3)$ pode ser facilmente estendido para o código $(8, 4, 4)$, bastando para tanto, acrescentar um bit de paridade extra conforme a Figura **??**.



Código de Hamming Estendido II

- Bits de dados: $x_0, x_1, x_2, x_3 \in \{0, 1\}$.
- Bits de paridade: $x_4, x_5, x_6, x_7 \in \{0, 1\}$.
- Os bits de paridade são determinados pelas equações:

$$x_4 = (x_1 + x_2 + x_3) \mod 2 \tag{763}$$

$$x_5 = (x_0 + x_2 + x_3) \mod 2 \tag{764}$$

$$x_6 = (x_0 + x_1 + x_3) \mod 2 \tag{765}$$

$$\begin{aligned} x_7 &= (x_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6) \mod 2 && (766)\\ &= (3x_0 + 3x_1 + 3x_2 + 4x_3) \mod 2 && (767)\\ &= (x_0 + x_1 + x_2) \mod 2 && (768) \end{aligned}$$

Código de Hamming Estendido III

- A matriz geradora será

$$G = \left(\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hdashline 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{array}\right) \tag{769}$$

- A matriz de verificação de paridade será

$$H = \left(\begin{array}{cccccccc} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ \hdashline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{array}\right), \tag{770}$$

Código de Hamming Estendido IV

ou, de forma equivalente,

$$H = \left(\begin{array}{cccccccc} 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ \hdashline 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right). \tag{771}$$

- A distância mínima do código de Hamming estendido $(8, 4, 4)$ é de 4. O código de Hamming $(7, 4, 3)$ possui distância mínima de 3.
- O número mínimo de 1s nas palavras no código de Hamming estendido $(8, 4, 4)$ é 4. Ou seja, o peso do código é 4. (Iremos verificar para a matriz dada na Equação 770)
  - Não podemos ter peso 1 pois as palavras com peso 1 não estão no espaço nulo de $\mathbf{H}$. Suponha que a palavra x seja não nula apenas na posição i. Então $\mathbf{H}x$ será igual à i-ésima coluna de $\mathbf{H}$. Mas não existe coluna nula em $\mathbf{H}$, desta forma não é possível ter $\mathbf{H}x = 0$ com palavras de peso 1.

Código de Hamming Estendido V

- Suponha que o peso do código seja 2, com elementos não nulos nas posições i e j. Neste caso, por ser uma palavra do código, devemos ter $\mathbf{H}x = 0$, assim a soma da i-ésima e j-ésima colunas de $\mathbf{H}$ deverá ser igual a zero. Como as colunas de $\mathbf{H}$ são todas diferentes, a soma de quaisquer duas colunas é não nula. Desta forma não é possível que o código tenha peso 2.
- Peso 3 também não é possível. Note que a última linha de $\mathbf{H}$ é toda igual a 1, logo ao somar 3 colunas de $\mathbf{H}$, no último índice estaremos somando 1 três vezes, e terá como resultado 1. Desta forma, será impossível obter como resultado o 0 desejado.
- Peso 4 será possível.

- A distância mínima entre palavras em C é 4.
  - Suponha que $v_1, v_2 \in C$ difiram em apenas uma posição. Sabemos que C é fechado em relação à soma, logo $(v_1 - v_2) \in C$. Como v_1, v_2 se diferem em apenas uma posição, $\mathbf{H}(v_1 - v_2)$ deverá ser uma coluna de $\mathbf{H}$, mas sabemos que não exite coluna de $\mathbf{H}$ nula. Não podemos ter v_1, v_2 diferindo em apenas uma posição.

Código de Hamming Estendido VI

- Suponha que $v_1, v_2 \in C$ difiram em apenas duas posições. Neste caso, $\mathbf{H}(v_1 - v_2)$ será a diferença ou soma de duas colunas de $\mathbf{H}$. Como $v_1, v_2 \in C$ teremos $(v_1 - v_2) \in C$. Não poderemos ter a diferença ou soma de quaisquer duas colunas de $\mathbf{H}$ igual a zero, desta forma v_1, v_2 não podem diferir em apenas duas posições.
- Suponha que $v_1, v_2 \in C$ difiram em apenas três posições. Mais uma vez chegaremos em contradição, pois teremos a diferença ou soma de quaisquer três colunas de $\mathbf{H}$, que não poderá ser igual a zero devido à última linha de $\mathbf{H}$ ser toda igual a 1.

- Procedimento de decodificação de síndrome para um dado y recebido:
  1) Calcular a síndrome $s = \mathbf{H}y$.
  2) Se $s = (000)^T$, faça $z \leftarrow (00000000)^T$ e vá ao passo 4.
  3) Caso contrário, localize a única coluna de $\mathbf{H}$ igual a s, faça z um vetor cheio de zeros mas com 1 na posição correspondente.
  4) Faça $x \leftarrow y + z$
  5) saída: (x_0, x_1, x_2, x_3)
- O procedimento é o mesmo utilizado no código do Hamming $(7, 4, 3)$.

Código de Hamming Estendido VII

- Este procedimento é capaz de corrigir um único bit de erro e detectar dois erros, mas falha quando mais do que dois bits são trocados (bits de erro).

Algoritmo do Código de Hamming I

		1	10	11	100	101	110	111	1000	1001	1010	1011	1100	1101	1110	1111	10000	10001	10010	10011	10100
posição do bit		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
bits de dado codificados		p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}	p_{14}	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	p_{20}
Cobertura dos bits de paridade	p_1	X		X		X		X		X		X		X		X		X		X	
	p_2		X	X			X	X			X	X			X	X			X	X	
	p_4				X	X	X	X					X	X	X	X					X
	p_8								X	X	X	X	X	X	X	X					
	p_{16}																X	X	X	X	X

- p_1: o primeiro bit de paridade abarcará as posições em que sua representação na forma binária apresenta o bit 1 em seu primeiro bit, ou seja, a posição menos significativa. São elas: 1 (**1**), 3 (1**1**), 5 (10**1**), 7 (11**1**), 9 (100**1**), 11 (101**1**), etc.
- p_2: o segundo bit de paridade abarcará as posições em que sua representação na forma binária apresenta o bit 1 em seu segundo bit, ou seja, a segundo posição menos significativa. São elas: 2 (**1**0), 3 (**1**1), 6 (1**1**0), 7 (1**1**1), 10 (10**1**0), 11 (10**1**1), etc.

Algoritmo do Código de Hamming II

- p_3: o terceiro bit de paridade abarcará as posições em que sua representação na forma binária apresenta o bit 1 em seu terceiro bit, ou seja, a terceira posição menos significativa. São elas: 4 (**1**00), 5 (**1**01), 6 (**1**10), 7 (**1**11), 12 (1**1**00), 13 (1**1**01), 14 (1**1**10), 15 (1**1**11), etc.
- e assim por diante...

Se utilizarmos m bits de paridade, poderemos realizar a verificação dos bits de 1 a $2^m - 1$. Subtraindo os bits de paridade, teremos $2^m - 1 - m$ bits que poderão ser utilizados para dados. Variando m, teremos os seguintes códigos de Hamming:

bits de paridade	total de bits	bits de dados	nome	taxa
2	3	1	Hamming(3,1)	$1/3 \approx 0.333$
3	7	4	Hamming(7,4)	$4/7 \approx 0.571$
4	15	11	Hamming(15,11)	$11/15 \approx 0.733$
5	31	26	Hamming(31,26)	$26/31 \approx 0.839$

Utilizamos anteriormente a seguinte formulação: $x^T = (x_0, x_1, \ldots, x_6)$ e

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{772}$$

Podemos trocar o ordenamento dos bits, sem prejuízo algum, e desta forma deveremos também trocar o ordenamento das colunas da matriz $\mathbf{H}$.
Suponha que seja feita a seguinte permutação $(x_0, x_1, \ldots, x_6) \to (x_4, x_2, x_3, x_0, x_1, x_5, x_6)$. A nova matriz $\mathbf{H}$ será dada então por

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}. \tag{773}$$

A soma de quais quer duas equações de restrição de paridade fornece uma nova equação. Por exemplo, se somarmos as equações representadas pela 1a e 2a linha, teremos

$$x_4 + 2x_2 + 2x_3 + x_0 + x_1 + x_5 = x_4 + x_0 + x_1 + x_5, \tag{774}$$

que é uma nova equação equação de verificação de paridade, a qual poderemos adicionar à matriz $\mathbf{H}$:

$$H' = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{775}$$

Note que a quarta linha é a soma (módulo 2) das linhas 1 e 2. Note ainda que todas as linhas são deslocamentos cíclicos de uma outra linha. Podemos ainda descartar uma das verificações de paridade redundante, descartando, por exemplo, a primeira linha, obtendo

$$H'' = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{776}$$

Realizando o mesmo procedimento acima para gerar as equações de paridade redundantes, podemos obter uma matriz super-redundante com sete linhas de verificação de paridade:

$$H' = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}. \tag{777}$$

Esta matriz é uma matriz cíclica, em que cada linha é uma permutação cíclica da primeira linha. *Códigos Cíclicos* são códigos que utilizam uma matriz cíclica para verificação de paridade. As palavras de código de tais códigos também possuem a propriedade de serem cíclicos, ou seja, qualquer permutação cíclica de uma palavra de código é também uma palavra de código.

Entropia I

- Caso discreto

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{778}$$

- O mundo é contínuo, os canais são contínuos, ruído é contínuo.
- Precisamos de uma teoria que possa ser aplicada ao domínio contínuo.

Entropia Contínua/Diferencial I

- Seja X uma v.a. contínua com distribuição cumulativa

$$F(x) = \Pr(X \leq x) \tag{779}$$

e a função densidade é dada pela derivada da cumulativa

$$f(x) = \frac{d}{dx} F(x). \tag{780}$$

- O suporte é definido por $S = \{x : f(x) > 0\}$.

Definição (entropia diferencial $h(X)$)

$$h(X) = -\int_S f(x) \log f(x) dx \tag{781}$$

- Como a integral é sobre o suporte S, não precisamos nos preocupar com $\log 0$.

Entropia Contínua/Diferencial II

Exemplo

Dado $X \sim \mathcal{U}[0, a]$ com $a \in R^+$, teremos

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} dx = -\log \frac{1}{a} = \log a \tag{782}$$

- Dependendo do valor de a, podemos ter entropia com valor positivo ou negativo, e ainda não é limitada.

- Entropia pode ser interpretada como o expoente do 'volume' do conjunto típico. Por exemplo, $2^{nH(X)}$ é o número de eventos que ocorrem, em média; podemos ter $2^{H(X)} \ll |\mathcal{X}|$. Teremos igualdade no caso uniforme. A incerteza de uma v.a. X é equivalente à de uma v.a. uniforme Y tal que $2^{H(X)} = |\mathcal{Y}|$.
- Expoente negativo significa que o 'volume' é pequeno.

Entropia Contínua/Diferencial III

Exemplo

Distribuição Normal (Gaussiana). Seja a v.a. $X \sim \mathcal{N}(0, \sigma^2)$, densidade dada por

$$f(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{1}{2}x^2/\sigma^2} \tag{783}$$

- Intuitivamente, sabemos que a entropia não depende da média μ.

...

Entropia Contínua/Diferencial IV

Exemplo

Entropia Contínua/Diferencial V

continuação...

- Calculando a entropia em nats

$$
\begin{aligned}
h(X) &= -\int f(x)\ln f(x)dx && (784)\\
&= -\int f(x)\left[-\frac{x^2}{2\sigma^2}-\ln\sqrt{2\pi\sigma^2}\right]dx && (785)\\
&= \frac{\mathrm{E}X^2}{2\sigma^2}+\frac{1}{2}\ln(2\pi\sigma^2)=\frac{1}{2}+\frac{1}{2}\ln(2\pi\sigma^2) && (786)\\
&= \frac{1}{2}\ln e+\frac{1}{2}\ln(2\pi\sigma^2)=\frac{1}{2}\ln(2\pi e\sigma^2)\text{nats} && (787)\\
&= \frac{1}{2}\ln(2\pi e\sigma^2)\text{nats}\times\left(\frac{1}{\ln 2}\text{bits/nats}\right)=\frac{1}{2}\log(2\pi e\sigma^2)\text{bits.}
\end{aligned}
$$

Propriedade da Equipartição Assintótica I

- No caso discreto: $\Pr(x_1, x_2, \ldots, x_n) \approx 2^{-nH(X)}$ para n grande suficiente e $|A_\epsilon^{(n)}| = 2^{nH} = (2^H)^n$.
- Desta forma, 2^H pode ser visto como o 'comprimento do lado' de um hipercubo em um espaço n dimensional, e assim 2^{nH} seria o volume deste hipercubo (ou volume do conjunto típico).
- H negativo implicaria em um comprimento pequeno 2^H, mas ainda positivo.

Propriedade da Equipartição Assintótica II

Teorema

Seja $X_1, X_2, \ldots, X_n$ uma sequência de v.a.s, i.i.d. $\sim f(x)$, então

$$-\frac{1}{n}\log f(X_1, \ldots, X_n) \to \mathrm{E}\left[-\log f(X)\right] = h(X) \tag{788}$$

Definição (Conjunto Típico)

$$A_\epsilon^{(n)} = \left\{ x_{1:n} \in S^n : \left| -\frac{1}{n}\log f(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \right\} \tag{789}$$

- Note que

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i). \tag{790}$$

Propriedade da Equipartição Assintótica III

- Limites para esta probabilidade

$$2^{-n(h+\epsilon)} \leq f(x_{1:n}) \leq 2^{-n(h-\epsilon)} \tag{791}$$

- O volume de $A \subseteq \mathbb{R}^n$ é definido

$$\text{Vol}(A) = \int_A dx_1 dx_2 \dots dx_n \tag{792}$$

Teorema

Propriedade da Equipartição Assintótica IV

1) *Probabilidade do conjunto típico*

$$\Pr(A_\epsilon^{(n)}) > 1 - \epsilon \tag{793}$$

2) *Volume do conjunto típico*

$$(1-\epsilon)2^{n(h(X)-\epsilon)} \leq \mathit{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)} \tag{794}$$

- Temos limites no volume do conjunto típico.
- No caso discreto, os limites eram sobre a cardinalidade do conjunto típico, e neste caso era necessário $H(X) \geq 0$, pois o tamanho mínimo de $|A_\epsilon^{(n)}|$ é 1.

Propriedade da Equipartição Assintótica V

Demonstração.

Por definição

$$\begin{aligned} p(A_\epsilon^{(n)}) &= \int_{x_{1:n} \in A_\epsilon^{(n)}} f(x_1, \ldots, x_n) dx_1 \ldots dx_n \\ &= \Pr\left(|-\frac{1}{n} f(x_1, x_2, \ldots, x_n) - h(X)| \leq \epsilon \right) \geq 1 - \epsilon \end{aligned} \tag{795}$$

para n grande suficiente, o que segue da lei fraca dos grandes números.

...

Propriedade da Equipartição Assintótica VI

Demonstração.

continuação...

Temos também que

$$\begin{aligned} 1 &= \int_{S^n} f(x_1,\ldots,x_n)dx_1\ldots dx_n \geq \int_{A_\epsilon^{(n)}} f(x_1,\ldots,x_n)dx_1\ldots dx_n \\ &\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)}dx_{1:n} = 2^{-n(h(X)+\epsilon)}\text{Vol}(A_\epsilon^{(n)}) \end{aligned} \tag{796}$$

Logo,

$$\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)} \tag{797}$$

...

Propriedade da Equipartição Assintótica VII

Demonstração.

continuação...

De forma similar,

$$1-\epsilon \leq \Pr(A_\epsilon^{(n)}) = \int_{A_\epsilon^{(n)}} f(x_{1:n})dx_{1:n} \tag{798}$$

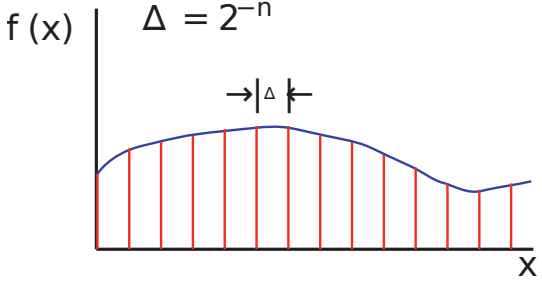
$$\leq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)}dx_{1:n} = 2^{-n(h(X)-\epsilon)}\text{Vol}(A_\epsilon^{(n)}) \tag{799}$$

□

- $A_\epsilon^{(n)}$ é o menor volume que contém toda a probabilidade e este volume é $\approx 2^{nh}$, e o tamanho do lado deste hipercubo é 2^h.
- Portanto, faz sentido $-\infty < h < \infty$.

Entropia Discreta vs Entropia Diferencial I

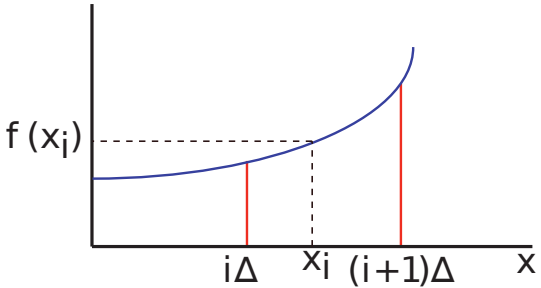
- Seja $X \sim f(x)$. Vamos dividir a extensão de X em *bins* (caixas) de tamanho Δ.
- Quantizar a extensão de X utilizando n bits, assim $\Delta = 2^{-n}$.



Entropia Discreta vs Entropia Diferencial II

- Pelo teorema do valor médio, se $f(\cdot)$ é contínua em um intervalo $[i\Delta, (i+1)\Delta]$, então $\exists x_i$ dentro deste intervalo, tal que

$$f(x_i) = \frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x)dx \tag{800}$$



Entropia Discreta vs Entropia Diferencial III

- Vamos criar uma variável aleatória quantizada X^{Δ} com os seguintes valores

$$X^{\Delta} = x_i \quad \text{se } i\Delta \leq X \leq (i+1)\Delta \tag{801}$$

- Obtemos assim a distribuição discreta

$$\Pr(X^{\Delta} = x_i) = p_i = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = \Delta f(x_i) \tag{802}$$

Entropia Discreta vs Entropia Diferencial IV

podemos calcular a entropia

$$H(X^{\Delta}) = -\sum_{i=-\infty}^{\infty} p_i \log p_i = -\sum_i f(x_i)\Delta \log(f(x_i)\Delta) \quad (803)$$

$$= -\sum_i f(x_i)\Delta \log f(x_i) - \sum_i f(x_i)\Delta \log \Delta \quad (804)$$

$$= -\sum_i f(x_i)\Delta \log f(x_i) - \log \Delta \sum_i \underbrace{f(x_i)\Delta}_{p_i} \quad (805)$$

$$= -\sum_i f(x_i)\Delta \log f(x_i) - \log \Delta \quad (806)$$

- Utilizamos que

$$\sum_i f(x_i)\Delta = \sum_i \Delta \left(\frac{1}{\Delta} \int_{i\Delta}^{(i+1)\Delta} f(x)dx \right) = \Delta \frac{1}{\Delta} \int f(x)dx = 1 \quad (807)$$

Entropia Discreta vs Entropia Diferencial V

- Quando $\Delta \to 0$, teremos que $-\log \Delta \to \infty$ e assim (assumindo que é integrável no sentido de Riemann)

$$-\sum_i \Delta f(x_i) \log f(x_i) \to -\int f(x) \log f(x) dx \tag{808}$$

- Desta forma,

$$H(X^\Delta) + \log \Delta \to h(f) \quad \text{quando } \Delta \to 0. \tag{809}$$

- De forma relaxada, podemos dizer que $h(f) \approx H(X^\Delta) + \log \Delta$ e para um quantizador de n bits com $\Delta = 2^{-n}$, temos

$$H(X^\Delta) \approx h(f) - \log \Delta = h(f) + n \tag{810}$$

- Isto significa que, quando $n \to \infty$, $H(X^\Delta)$ torna-se maior.
- Começamos com uma v.a. contínua X e quantizamos com uma acerácea de n bits. Para uma representação discreta com 2^n valores, esperamos que a entropia cresça com n.

Entropia Discreta vs Entropia Diferencial VI

- $H(X^{\Delta})$ é o número de bits necessários para descrever esta quantização uniforme em n bits da v.a. X.
- $H(X^{\Delta}) \approx h(f) + n$, então podemos precisar de mais ou menos do que n bits para descrever X com uma acerácea de n bits, dependendo da concentração de X.
- Se X for muito concentrado $h(f) < 0$ e desta forma precisaremos de menos do que n bits. Se X for muito espalhado, precisaremos de mais do que n bits.

Entropia Diferencial Conjunta I

Definição (Entropia Diferencial Conjunta)

$$h(X_1, X_2, \ldots, X_n) = -\int f(x_{1:n}) \log f(x_{1:n}) dx_{1:n} \tag{811}$$

Definição (Entropia Diferencial Condicional)

$$h(X \mid Y) = -\int f(x, y) \log f(x \mid y) dx dy = h(X, Y) - h(Y) \tag{812}$$

Entropia da Gaussiana Multidimensional I

- $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, possui distribuição dada por uma Gaussiana multivariável, dada pelo vetor de média μ e a matriz de covariância $\mathbf{\Sigma}$, ou seja,

$$f(\mathbf{x}) = \frac{1}{|2\pi\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)} \tag{813}$$

  onde $\mathbf{x} = x_{1:n}$.

- A entropia de $\mathbf{X}$ será dada por

$$h(\mathbf{X}) = \frac{1}{2} \log \left[(2\pi e)^n |\mathbf{\Sigma}|\right] \text{ bits.} \tag{814}$$

  onde $|\mathbf{\Sigma}|$ é o determinante da matriz de covariância.

- Note que a entropia é monotonicamente relacionada com o determinante da matriz de covariância $\mathbf{\Sigma}$ e não depende da média μ.

Entropia da Gaussiana Multidimensional II

- Reescrevendo a equação acima, podemos obter $|\mathbf{\Sigma}|$ como uma função da entropia, e desta forma teremos $|\mathbf{\Sigma}| \propto 2^{h(\mathbf{X})}$.
- O determinante da matriz de covariância é uma medida de dispersão (espalhamento) da distribuição.
- Para encontrar a entropia, faremos

$$
\begin{aligned}
h(\mathbf{X}) &= -\int f(x) \ln f(x) dx && (815)\\
&= -\int f(x) \left[-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) - \ln\left((2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}\right) \right] dx \\
&= \frac{1}{2}\int f(x)(\mathbf{x}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu) dx \\
&\quad + \ln\left((2\pi)^{n/2}|\mathbf{\Sigma}|^{1/2}\right) \int f(x) dx && (816)\\
&= \frac{1}{2}\mathrm{E}_f\left[\mathrm{Tr}(\mathbf{X}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)\right] + \frac{1}{2}\ln[(2\pi)^n|\mathbf{\Sigma}|] && (817)
\end{aligned}
$$

Entropia da Gaussiana Multidimensional III

- Iremos utilizar a seguinte propriedade do traço:

$$\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB) \tag{818}$$

Entropia da Gaussiana Multidimensional IV

- continuando

$$
\begin{aligned}
h(\mathbf{X}) &= \dots \\
&= \frac{1}{2}\mathrm{E}_f\left[\mathrm{Tr}(\mathbf{X}-\mu)^T\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mu)\right] + \frac{1}{2}\ln[(2\pi)^n|\boldsymbol{\Sigma}|] && (819)\\
&= \frac{1}{2}\mathrm{E}_f\left[\mathrm{Tr}(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T\boldsymbol{\Sigma}^{-1}\right] + \frac{1}{2}\ln[(2\pi)^n|\boldsymbol{\Sigma}|] && (820)\\
&= \frac{1}{2}\,\mathrm{Tr}\,\mathrm{E}_f\left[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^T\right]\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\ln[(2\pi)^n|\boldsymbol{\Sigma}|] && (821)\\
&= \frac{1}{2}\,\mathrm{Tr}\,\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\ln[(2\pi)^n|\boldsymbol{\Sigma}|] && (822)\\
&= \frac{1}{2}\,\mathrm{Tr}\,\mathbf{I} + \frac{1}{2}\ln[(2\pi)^n|\boldsymbol{\Sigma}|] = \frac{1}{n} + \frac{1}{2}\ln[(2\pi)^n|\boldsymbol{\Sigma}|] && (823)\\
&= \frac{1}{2}\ln[(2\pi e)^n|\boldsymbol{\Sigma}|] && (824)
\end{aligned}
$$

Entropia Relativa / Divergência de KL I

Definição (Entropia Relativa / Divergência de KL)

$$D(f \mid\mid g) = \int f(x) \log \frac{f(x)}{g(x)} dx \geq 0 \tag{825}$$

- Podemos utilizar a desigualdade de Jensen para provar a não-negatividade de $D(f \mid\mid g)$.

Definição (Informação Mútua)

$$\begin{aligned} D(f(X,Y) \mid\mid f(X)f(Y)) &= I(X;Y) = h(X) - h(X \mid Y) && (826) \\ &= h(Y) - h(Y \mid X) \geq 0 && (827) \end{aligned}$$

Entropia Relativa / Divergência de KL II

- Como $I(X;Y) \geq 0$, temos novamente que condicionar reduz entropia, i.e., $h(Y) \geq h(Y \mid X)$.

Regra da Cadeia e outros I

- Regra da Cadeia

$$h(X_1, X_2, \ldots, X_n) = \sum_i h(X_i \mid X_{1:i-1}) \tag{828}$$

- Limites

$$\sum_i h(X_i \mid X_{1:n\setminus\{i\}}) \leq h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i) \tag{829}$$

- Para entropia discreta temos a monotonicidade, i.e.,

$$H(X_1, X_2, \ldots, X_k) \leq H(X_1, X_2, \ldots, X_k, X_{k+1}). \tag{830}$$

De forma geral,

$$f(A) = H(X_A) \tag{831}$$

é monotônica não-decrescente no conjunto A (i.e., $f(A) \leq f(B), \forall A \subseteq B$).

Regra da Cadeia e outros II

- No caso contínuo, teremos que $f(A) = h(X_A)$ não é monotônico. Considere o exemplo da Gaussiana com diagonal de Σ com valores pequenos. Então $h(X) = \frac{1}{2}\log\left[(2\pi e)^n|\Sigma|\right]$ pode ficar menor com mais variáveis aleatórias.
- De forma similar, quando temos v.a.s independentes, adicionar aquelas que possuem entropia negativa pode diminuir a entropia total.

Translação I

Teorema

A translação não afeta a entropia diferencial, ou seja,

$$h(X+c) = h(X). \tag{832}$$

Demonstração.

$$\begin{aligned} h(X+c) &= -\int_{S+c} f(x+c)\log f(x+c)dx & (833)\\ &= -\int_{S} f(x)\log f(x)dx = h(X) & (834)\end{aligned}$$

□

Mudança de Escala I

Teorema

Uma mudança de escala na variável independente acarreta na soma de uma constante à entropia, ou seja,

$$h(aX) = h(X) + \log|a|. \tag{835}$$

Mudança de Escala II

Demonstração.

Mudança de Escala III

Seja $Y = aX$, então $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y}{a}\right)$, e

$$\begin{aligned}
h(aX) &= -\int_{S_y} f_Y(y) \log f_Y(y) dy \\
&= -\int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|} f_X\left(\frac{y}{a}\right)\right) dy \\
&= -\int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(\frac{1}{|a|}\right) dy - \int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) \log\left(f_X\left(\frac{y}{a}\right)\right) dy \\
&= \log|a| \underbrace{\int_{S_y} \frac{1}{|a|} f_X\left(\frac{y}{a}\right) dy}_{=1} - \int_{S_x} f_X(x) \log f_X(x) dx \\
&= \log|a| + h(X) \qquad (836)
\end{aligned}$$

□

Corolário

Desigualdade de Hadamard I

- Como $h(X_1, X_2, \ldots, X_n) \leq \sum_i h(X_i)$, considere o caso em que $X_{1:n}$ é conjuntamente Gaussiano $\sim \mathcal{N}(\mu, K)$.
- Teremos então

$$\frac{1}{2}\log\left[(2\pi e)^n |K|\right] \leq \sum_i \frac{1}{2}\log\left[(2\pi e) K_{ii}\right] \quad (838)$$

$$(2\pi e)^n |K| \leq \prod_i (2\pi e) K_{ii} \quad (839)$$

$$|K| \leq \prod_i K_{ii} \quad (840)$$

onde utilizamos que o $\log$ é monotônico. K é positiva semi-definida, pois seu determinante é limitado pelo produto dos elementos na diagonal.

Entropia Máxima e Gaussianas I

Teorema

Uma Gaussiana possui entropia máxima dentre todas as distribuições que possuem os mesmos momentos de primeira e segunda ordem. Isto é, seja $X \in \mathbb{R}^n$ um vetor variável aleatória com $\mathrm{E}X = 0$ e $\mathrm{E}XX^T = K$, então

$$h(X) \leq \frac{1}{2} \log(2\pi e)^n |K| \tag{841}$$

com igualdade quando $X \sim \mathcal{N}(0, K)$.

Entropia Máxima e Gaussianas II

Demonstração.

- Seja $g(X)$ uma distribuição com média nula tal que o seu segundo momento seja igual à K, a matriz de covariância, ou seja,

$$\int g(x) X X^T dx = K \tag{842}$$

- Seja a distribuição gaussiana $\eta(X) \sim \mathcal{N}(0, K)$, então

$$\int \eta(x) X X^T dx = K \tag{843}$$

...

Entropia Máxima e Gaussianas III

Demonstração.

Entropia Máxima e Gaussianas IV

continuação...

- Observamos que $\log \eta(X)$ possui forma quadrática, i.e.,

$$\log \eta(x) = -\frac{1}{2} x^T K^{-1} x - \frac{1}{2} \ln[(2\pi)^n |K|] \tag{844}$$

- Teremos assim

$$\begin{aligned} 0 \quad &\leq \quad D(g \mid\mid \eta) = \int g(x) \log g(x)/\eta(x) dx && (845) \\ &= \quad -h(g(x)) - \int g(x) \log \eta(x) dx && (846) \\ &= \quad -h(g(x)) + \int g(x) \left(\frac{1}{2} x^T K^{-1} x + \frac{1}{2} \ln[(2\pi)^n |K|] \right) && (847) \end{aligned}$$

...

Entropia Máxima e Gaussianas V

Demonstração.

continuação...

$$
\begin{aligned}
0 &\leq \dots \\
&= -h(g(x)) + \frac{1}{2}\int g(x)[x^T K^{-1} x]dx + \frac{1}{2}\ln[(2\pi)^n |K|] && (848)\\
&= -h(g(x)) + \frac{1}{2}\mathrm{E}_g[\mathrm{Tr}(X^T K^{-1} X)] + \frac{1}{2}\ln[(2\pi)^n |K|] && (849)\\
&= -h(g(x)) + \frac{1}{2}\mathrm{E}_g[\mathrm{Tr}(K^{-1} X X^T)] + \frac{1}{2}\ln[(2\pi)^n |K|] && (850)\\
&= -h(g(x)) + \frac{1}{2}\mathrm{Tr}(K^{-1}\mathrm{E}_g[X X^T]) + \frac{1}{2}\ln[(2\pi)^n |K|] && (851)\\
&= -h(g(x)) + \frac{1}{2}\mathrm{Tr}(K^{-1} K) + \frac{1}{2}\ln[(2\pi)^n |K|] && (852)
\end{aligned}
$$

...

Entropia Máxima e Gaussianas VI

Demonstração.

continuação...

$$
\begin{aligned}
0 &\leq \dots \\
&= -h(g(x)) + \frac{1}{2}\operatorname{Tr}(I) + \frac{1}{2}\ln[(2\pi)^n|K|] && (853)\\
&= -h(g(x)) + \frac{1}{n} + \frac{1}{2}\ln[(2\pi)^n|K|] && (854)\\
&= -h(g(x)) + \frac{1}{2}\ln[(2\pi e)^n|K|] && (855)\\
&= -h(g(x)) + h(\eta(x)) && (856)
\end{aligned}
$$

Logo, teremos que $h(\eta(x)) \geq h(g(x))$.

□

Erro de Estimação e Entropia Diferencial I

Teorema

Para qualquer variável aleatória X e estimador $\hat{X}$,

$$\mathrm{E}\left(X-\hat{X}\right)^2 \geq \frac{1}{2\pi e} e^{2h(X)}, \tag{857}$$

com igualdade sse X é Gaussiana e $\hat{X}$ é a média de X.

Erro de Estimação e Entropia Diferencial II

Demonstração.

Seja $\hat{X}$ um estimador de X, então

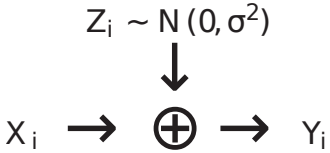
$$\begin{aligned}
\mathrm{E}\left(X-\hat{X}\right)^2 &\geq \min_{\hat{X}} \mathrm{E}\left(X-\hat{X}\right) && (858)\\
&\quad \text{a média é o melhor estimador para } X\\
&= \mathrm{E}\left(X-\mathrm{E}(X)\right)^2 && (859)\\
&= \mathrm{Var}(X) && (860)\\
&\quad \text{a distribuição gaussiana possui entropia máxima}\\
&\geq \frac{1}{2\pi e} e^{2h(X)} && (861)
\end{aligned}$$

□

Canais Contínuos I

- Foi visto até o momento apenas canais discretos, modelados pela distribuição de probabilidade condicional $p(y|x)$.
- De forma geral os canais de comunicação são contínuos e os sinais reais. O que ocorre em verdade é que, dada uma v.a. X, teremos $Y = Z(X)$, onde Z é uma função aleatória que pode depender ou não de X.
- Isto é difícil de se analisar, portanto iremos considerar um modelo mais simples: ruído aditivo onde $Y = X + Z$, sendo Z uma v.a..
- Podemos simplificar ainda mais assumindo $Z \perp\!\!\!\perp X$.
- Podemos considerar Z Gaussiano.

Canal Gaussiano I



- Modelo em que $Y_i = X_i + Z_i$ com $Z_i \sim \mathcal{N}(0, \sigma^2)$ e $Z_i \perp\!\!\!\perp X_i$.
- Quando $\sigma^2 = 0$ teremos um canal com capacidade infinita, pois será possível enviar um número real qualquer com precisão, o que pode requerer um número infinito de bits (obs: na codificação aritmética utilizamos uma sequência de bits tão longa quanto necessária para codificar um número no intervalo $[0, 1)$).
- Se $\sigma^2 > 0$ podemos ter capacidade ainda infinita, pois poderemos aumentar a potência do sinal, tornando o ruído insignificante.
- Iremos tratar do problema mais realístico, quando a potência é limitada.

Limitação de Potência I

- Restrição média sobre a potência: para qualquer palavra de comprimento n, vamos impor a seguinte restrição:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P \tag{862}$$

  onde P é a potência média $\approx \mathrm{E}X^2$.

- Outras restrições também são possíveis, por exemplo: limitar o valor máximo; limitar o valor máximo em uma determinada janela.

Exemplo I

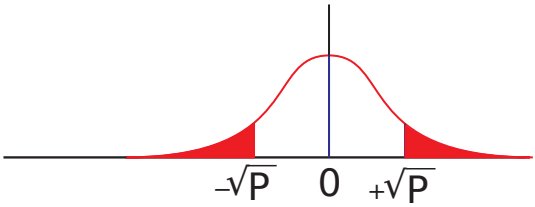
- Enviar 1 bit por vez através do canal.
- $X \in \{-\sqrt{P}, +\sqrt{P}\}$, desta forma, $\mathrm{E}X^2 = P$, satisfazendo a restrição.
- Para X com uma distribuição uniforme iremos decodificar como $+\sqrt{P}$ quando $Y > 0$ e como $-\sqrt{P}$ quando $Y < 0$.
- Erro:

$$
\begin{aligned}
P_e &= \Pr(Y < 0 \mid X = +\sqrt{P}) \Pr(X = +\sqrt{P}) + \\
&\quad \Pr(Y > 0 \mid X = -\sqrt{P}) \Pr(X = -\sqrt{P}) && (863) \\
&= \Pr(Z < -\sqrt{P} \mid X = +\sqrt{P}) \frac{1}{2} + \Pr(Z > +\sqrt{P} \mid X = -\sqrt{P}) \frac{1}{2} && (864) \\
&\quad \text{como } Z \perp\!\!\!\perp X \\
&= \Pr(Z > \sqrt{P}) && (865)
\end{aligned}
$$

- Z possui distribuição Gaussiana. Os dois tipos de erro (quando $Y < 0 \mid X = +\sqrt{P}$ e $Y > 0 \mid X = -\sqrt{P}$) são ilustrados abaixo.

Exemplo II

- Erro total ($\times 1/2$):



Exemplo III

- Teremos assim

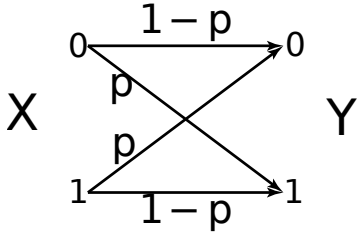
$$\Pr(Z > \sqrt{P}) = 1 - \Phi\left(\frac{\sqrt{P}}{\sigma^2}\right) \tag{866}$$

onde Φ é a distribuição normal cumulativa, i.e.,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \tag{867}$$

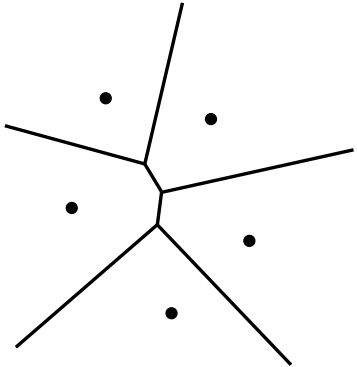
- Desta forma, transformamos um canal Gaussiano em um canal binário simétrico onde $p = P_e$.

Exemplo IV



- Podemos converter um canal contínuo em um canal discreto, utilizando uma codificação apropriada para tanto.
- Este é um processo de quantização vetorial. Para cada esquema de quantização devemos analisar a relação de compromisso entre a taxa e a distorção sob determinada restrição de potência.

Exemplo V



Capacidade do Canal Gaussiano I

Definição

A capacidade (de informação com restrição de potência P) é definida como

$$C = \max_{p(x):\mathrm{E}X^2 \leq P} I(X;Y) \text{bits} \tag{868}$$

- Aqui foi dada apenas a definição, nada foi dito com relação à possibilidade de comunicar a uma taxa igual, menor ou maior a esta capacidade.
- $I(X;Y)$ será da seguinte forma

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(X+Z|X) && (869) \\ &= h(Y) - h(Z|X) = h(Y) - h(Z) && (870) \end{aligned}$$

- Note que, quando $X \perp\!\!\!\perp Z$, teremos $h(X+Z) \geq h(Z)$, e desta forma teremos $I(X;Y) \geq 0$.

Capacidade do Canal Gaussiano II

- Estratégia para encontrar C: 1) limitar $I(X;Y)$; 2) encontrar $p(x)$ (não necessariamente única) que alcança o limite.
- Z é Gaussiana, logo $h(Z) = \frac{1}{2}\log(2\pi e\sigma^2)$ onde σ^2 é a potência do ruído, $\mathrm{E}Z^2 = \sigma^2 = N$, com $\mathrm{E}Z = 0$.
- Vimos anteriormente que a entropia de uma Gaussiana é limitada pelo segundo momento da seguinte forma, considerando $\mathrm{E}X = 0$ e $\mathrm{Var}X = K$,

$$h(X) \leq \frac{1}{2}\log[(2\pi e)^2|K|] \tag{871}$$

- Teremos também

$$\mathrm{E}Y^2 = \mathrm{E}(X+Z)^2 = \mathrm{E}X^2 + \underbrace{2\mathrm{E}X\mathrm{E}Z}_{\text{pois } X \perp\!\!\!\perp Z} + \mathrm{E}Z^2 \tag{872}$$

$$= \underbrace{P}_{\text{potência do sinal, E}X^2} + \underbrace{\sigma^2}_{\text{potência do ruído, E}Z^2} \tag{873}$$

Capacidade do Canal Gaussiano III

- Podemos limitar a informação mútua, utilizando o limite para a entropia de uma Gaussiana.

$$\begin{aligned} I(X;Y) &= h(Y) - h(Z) \leq \frac{1}{2}\log(2\pi e(P+\sigma^2)) - \frac{1}{2}\log(2\pi e\sigma^2) \\ &= \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right) = \frac{1}{2}\log(1+\text{SNR}) \end{aligned} \tag{874}$$

  onde SNR é a relação sinal-ruído.
- Como visto anteriormente, a Gaussiana possui entropia máxima dentre todas as distribuições que possuem os mesmos momentos de primeira e segunda ordem.
- Poderemos alcançar a informação mútua máxima se garantirmos que Y é gaussiano, ou seja, como Z *a priori* gaussiano, devemos escolher X gaussiano, pois a soma de gaussianas é gaussiana, e assim garantiremos que Y é gaussiano.

Capacidade do Canal Gaussiano IV

- Desta forma, a capacidade de um canal Gaussiano é

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) = \frac{1}{2}\log(1 + \mathsf{SNR}) \tag{875}$$

- A capacidade será alcançada quando $X \sim \mathcal{N}(0, P)$.

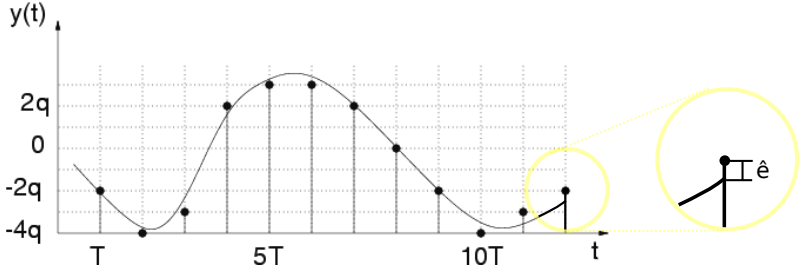
Exemplo: PCM 6dB SNR/bit I

- Dado um sinal limitado em frequência $x(t)$ tal que $X(f) = 0$ para $f > f_c$, se amostrarmos este sinal a uma taxa $1/\tau \geq 2f_c$ (taxa de Shannon/Nyquist), teremos reconstrução perfeita.

$$x[n] = x(n\tau) \tag{876}$$

  onde τ é o período de amostragem.

- Cada uma das amostras é quantizada e representada por um inteiro.



Exemplo: PCM 6dB SNR/bit II

- A quantização pode ser modelada por um canal gaussiano, i.e.,

$$\hat{x}[n] = x[n] + \eta_n = x(n\tau) + \eta_n \tag{877}$$

onde $\hat{x}[n] = jq$ para um inteiro $q \in \mathbb{Z}$, q é o quantum (passo) de quantização e $\eta_n \sim \mathcal{N}(0, \sigma^2)$ é um ruído gaussiano aditivo.

- Utilizando b bits para representar cada valor quantizado, teremos no máximo 2^b valores possíveis diferentes para $\hat{x}[n]$.
- Utilizando a equação obtida para a capacidade de uma canal Gaussiano

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) = \frac{1}{2}\log(1 + \text{SNR}) \tag{878}$$

- Para 6dB que aumentamos a SNR, será necessário $\frac{1}{2}\log(1+2)$ bits a mais na capacidade do canal.
- Tipicamente, para áudio, temos $b = 16$ bits e, desta forma, SNR de 96 dB.

Exemplo: PCM 6dB SNR/bit III

- Teremos assim

$$16\text{bits/utilização do canal} = \frac{1}{2}\log(1+\text{SNR}) \tag{879}$$

desta forma, $2^{32} = 1 + \text{SNR}$ e assim $\text{SNR} = 2^{32} - 1$.

- Em escala logaritmica

$$10\log_{10}(\text{SNR}) \approx 10 \times 32 \log_{10} 2 \approx 96.33\text{dB} \tag{880}$$

e assim teremos $96.33/16 \approx 6.02$dB/bit.

- Cada bit adicional em um áudio PCM adiciona 6.02 dB na SNR.

Capacidade de Canal I

Definição (código)

Um código (M, n) para um canal Gaussiano, com restrição de potência P, inclui

1) conjunto de índices $\{1, 2, \ldots, M\}$
2) função de codificação $X : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, fornecendo palavras $X^n(1), X^n(2), \ldots, X^n(M)$ tais que

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2(\omega) \leq P, \quad \forall \omega \in \{1, \ldots, M\} \tag{881}$$

3) função de decodificação

$$g : \mathcal{Y}^n \to \{1, \ldots, M\} \tag{882}$$

Capacidade de Canal II

Definição (taxa)

A taxa é dada por

$$R = \frac{\log M}{n} \text{ bits por utilização do canal} \tag{883}$$

Definição (taxa alcançável)

Um taxa R é alcançável se $\exists$ uma sequência de códigos $(2^{nR}, n)$ satisfazendo a restrição de potência P tal que $\lambda^{(n)} \to 0$ (a probabilidade máxima de erro) quando $n \to \infty$.

Definição (capacidade)

A capacidade de um canal Gaussiano é o supremo de todas as taxas alcançáveis (i.e., o maior valor possível alcançável).

Capacidade de Canal III

Teorema (Capacidade do canal Gaussiano)

A capacidade de um canal Gaussianoo com restrição P na potência de entrada e ruído com variância σ^2 é dada por

$$C = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) \text{ bits por utilização do canal} \tag{884}$$

Demonstração.

- O conjunto típico em X, $A_\epsilon^{(n)}$, possui volume $\leq 2^{n(h(X)+\epsilon)}$.
- O conjunto condicionalmente típico em Y possui volume $\leq 2^{n(h(Y|X)+\epsilon)} = 2^{n(h(Z)+\epsilon)}$.
- O conjunto típico em Y (incondicional) possui volume $\leq 2^{n(h(Y)+\epsilon)}$, mas temos que $h(Y) \leq \frac{1}{2}\log[2\pi e(P+\sigma^2)]$ e $h(Z) = \frac{1}{2}\log[2\pi e\sigma^2]$.

...

Capacidade de Canal IV

Demonstração.

continuação...

- Quantos volumes condicionais em X podemos empacotar dentro do volume total?

$$\leq \frac{2^{nh(Y)}}{2^{nh(Z)}} = \frac{2^{n\frac{1}{2}\log[2\pi e(P+\sigma^2)]}}{2^{n\frac{1}{2}\log[2\pi e\sigma^2]}} \approx 2^{\frac{n}{2}\log\frac{P+\sigma^2}{\sigma^2}} = [(P+\sigma^2)/\sigma^2]^{n/2} \tag{885}$$

- Iremos assumir o melhor cenário (taxa máxima), sem sobreposição dos volumes.
- Teremos

$$2^{nR} = 2^{\frac{n}{2}\log\frac{P+\sigma^2}{\sigma^2}} \text{ e assim, } R = \frac{1}{2}\log(1+P/\sigma^2) \tag{886}$$

...

Capacidade de Canal V

Demonstração.

continuação...

- Se tudo for conjuntamente Gaussiano i.i.d., os volumes típicos serão esferas.
- Volume da esfera (em n-D)

$$V(r,n) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} r^n = 2^{\frac{n}{2}\log[2\pi e\sigma^2]} = (2\pi e r^2)^{n/2} \tag{887}$$

teremos assim

$$r_{\sigma^2} = \Gamma^{1/2}\left(\frac{n}{2}+1\right)(2e\sigma^2)^{1/2} \tag{888}$$

...

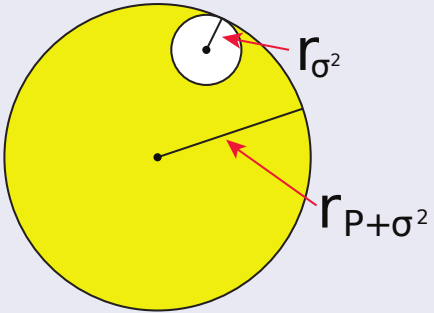
Capacidade de Canal VI

Demonstração.

Capacidade de Canal VII

continuação...

- Quantas esferas pequenas cabem dentro da esfera maior?



- Semelhente ao que foi feito no caso discreto (a razão entre os volumes fornece o limite).

...

Capacidade de Canal VIII

Demonstração.

continuação...

- Precisamos mostrar que se $R < C$, $\exists$ um código com $P_e^n \to 0$ quando $n \to \infty$.
- Geramos palavras de código aleatórias (assim como feito no caso discreto), mas neste caso com Gaussianas com $\mathrm{E}X^2 = P - \epsilon$, de forma que

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \to P - \epsilon \text{ quando } n \to \infty \tag{889}$$

...

Capacidade de Canal IX

Demonstração.

continuação...

- Teremos uma fonte adicional de erro possível (quando não for satisfeita a restrição em potência)

$$E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i^2(1) > P \right\} \tag{890}$$

- Devemos adicionar E_0 aos demais erros vistos anteriormente no caso discreto.
- Pela lei fraca dos grandes números, $E_0 \to 0$ quando $n \to \infty$ assim como para as demais fontes de erro.

□

Limitação em Banda I

- Suponha que o canal seja limitado em frequência, ou seja, $H(j\omega) = 0$ para $|\omega| > W$, onde W é a largura de banda do canal, ω é a variável que representa frequência e $H(j\omega)$ é a resposta em frequência do canal, ou seja, transformada de Fourier de $h(t)$ (resposta ao impulso).
- Podemos ver a saída como

$$Y(t) = (X(t) + Z(t)) * h(t) \tag{891}$$

  desta forma, se o canal for limitado em frequência, teremos um sistema não-causal, ele agirá como enviando a saída através de um filtro passa-baixas antes de observá-la.
- Resposta ao impulso de um filtro limitado em banda.

$$h(t) \overset{\mathsf{FT}}{\longleftrightarrow} H(j\omega) \tag{892}$$

- Como esta restrição irá influenciar a capacidade?

Limitação em Banda II

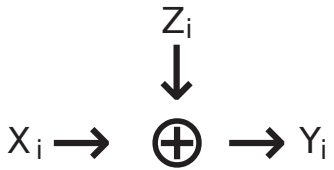
- Se o sinal é limitado em banda com limite W, quantas amostras por segundo serão necessárias? $2W$ amostras por segundo, de forma que teremos uma amostra a cada $\tau = 1/2W$ segundos.
- Grosseiramente, se um sinal é 'aproximadamente' limitado em frequência (W) e no tempo (T), podemos descrever o sinal apropriadamente com $2WT$ coeficientes/amostra.
- Dado um sinal $x(t)$ aproximadamente limitado em tempo e frequência, utilizando uma base apropriada (por exemplo, senoides truncadas de comprimento T), a representação

$$x(t) = \sum_{i=-WT}^{WT} \alpha_i \psi_i(t) + x_r(t) \tag{893}$$

terá erro residual pequeno, i.e., $\|x_r(t)\|$ é pequeno quando utilizarmos $2WT$ coeficientes $\{\alpha_i\}_i$.

- Modelo do ruido aditivo.

Limitação em Banda III



- Vamos assumir que o ruído possui densidade espectral de potência $\sigma^2/2$ watts/hertz e largura de banda de W hertz, então a potência do ruído é $\frac{\sigma^2}{2}2W = \sigma^2 W$.
- A energia total é $\sigma^2 WT$ (potência = energia por unidade de tempo ou, neste caso, por amostra).
- Cada amostra possui variância: energia total / número de amostras $= \sigma^2 WT/(2WT) = \sigma^2/2$.
- $Z = Y|X$ possui distribuição da forma

$$Y|X \sim \mathcal{N}\left(0, \frac{\sigma^2}{2} I\right) \tag{894}$$

Limitação em Banda IV

- Isto pode ser visto como uma esfera típica no espaço de recepção.
- A potência do sinal por amostra da entrada dada por: energia total / número de amostras $= TP/2WT = P/2W$.
- A potência do ruído por amostra da saída é $\sigma^2/2$.
- Utilizando esses valores na fórmula da capacidade em função da SNR = potência do sinal por amostra / variância do ruído por amostra, teremos

$$C = \frac{1}{2}\log\left(1 + \frac{P/2W}{\sigma^2/2}\right) = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2 W}\right) \text{ bits/amostra} \tag{895}$$

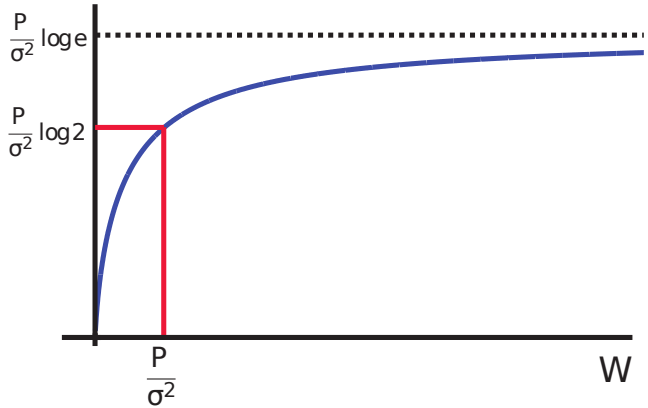
- Como existem $2W$ amostras por segundo, teremos

$$C = W\log\left(1 + \frac{P}{\sigma^2 W}\right) \text{ bits por segundo} \tag{896}$$

- Podemos aumentar a capacidade: 1) aumentando a largura de banda W; 2) aumentando a potência do sinal P ou; 3) reduzindo a variância do ruído σ^2.

Limitação em Banda V

- Gráfico de C em função de W para σ^2 fixo.



Limitação em Banda VI

- Observamos que C cresce rápidamente quando $W \in [0, P/\sigma^2]$ e a partir deste ponto C cresce mais devagar com $C_\infty = \left(\frac{P}{\sigma^2}\right) \log e$.

Exercício (Linha Telefônica)

Para permitir a multiplexação de diversos canais em uma linha telefônica, o sinal telefônico é limitado a uma banda com largura de 3300 Hz. Supondo que a linha telefônica esteja sujeita a ruído e que a relação sinal-ruído máxima da linha é de 33 dB, encontre a capacidade do canal telefônico em bits por segundos. (Os modems reais atingem uma taxa de transmissão de até 33.600 bits por segundo. Em linhas telefônicas reais outros fatores influenciam a capacidade do canal como, por exemplo, interferências, linhas cruzadas, ecos, e característica não-plana do canal telefônico. ...

Limitação em Banda VII

Exercício (Linha Telefônica)

continuação...

Os modems V.90 atingiam 56 kb/s em apenas uma direção utilizando o canal telefônico, tirando proveito da linha digital pura entre o servidor e a central telefônica final da rede. Nesta caso, os únicos empecilhos são devidos a conversão analógico-digital na central e o ruído no link de cobre da central até a residência: estes empecilhos reduzem a taxa de bits máxima de 64 kb/s do sinal digital na rede para 56 kb/s nas melhores linhas telefônicas. A largura de banda efetiva do fio de cobre é da ordem de alguns megahertz; o que depende do comprimento da linha. A resposta em frequência também não é plana. Se toda a largura de banda for utilizada, é possível enviar alguns megabits por segundo através do canal; o que é o caso do DSL (Digital Subscriber Line). ...

Limitação em Banda VIII

Exercício (Linha Telefônica)

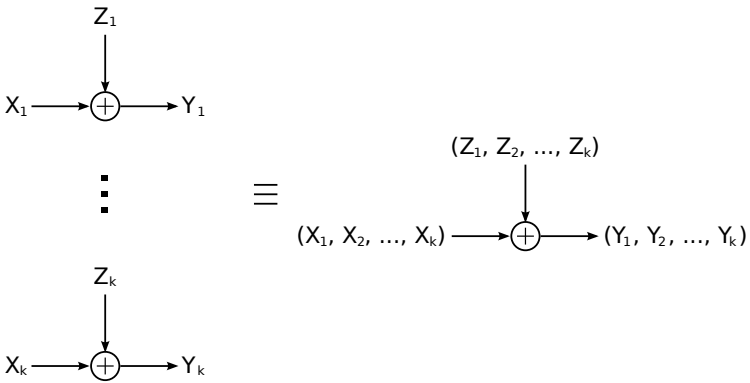
continuação...

(solução) Vamos utilizar $10^{3.3} \approx 2000$ *e* $\log_2 1000 = 9.9658$. *O ruído possui densidade de potência espectral* $N_0/2$ *watts/hertz e largura de banda* W. *A potência do ruído é* $\frac{N_0}{2} 2W = N_0 W$.

$$
\begin{aligned}
C &= W \log \left(1 + \frac{P}{N_0 W}\right) && (897)\\
&= W \log (1 + SNR) = 3300 \log(1 + 10^{3.3}) && (898)\\
&= 3300 \log (2000) = 3300(\log 2 + \log 1000) && (899)\\
&= 3300 \times (1 + 9.9658) && (900)\\
&= 36187 \text{ bits / s} && (901)
\end{aligned}
$$

Canais em Paralelo I

- Sejam k canais em paralelo.



Canais em Paralelo II

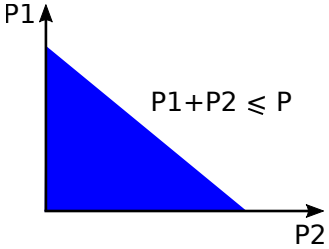
- O ruido é caracterizado por

$$\begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_k \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} N_1 & 0 & \dots & & 0 \\ 0 & N_2 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & & 0 & N_k \end{bmatrix}\right) \tag{902}$$

- Os ruídos não são correlacionados já que $Z_i \perp\!\!\!\perp Z_j$ para $i \neq j$.
- Sem restrições, a capacidade será $C = \log\left(\sum_i 2^{C_i}\right)$, se utilizarmos um canal por vez, e será $\sum_i C_i$, se utilizarmos simultâneamente.
- Qual será a capacidade se houve uma restrição comum de potência?

$$\mathrm{E}\left[\sum_{j=1}^{k} X_j^2\right] = \sum_{j=1}^{k} \mathrm{E}X_j = \sum_i P_i \leq P \tag{903}$$

Canais em Paralelo III



- Objetivo agora é encontrar a capacidade, dada a restrição de potência.

$$C = \max_{f(x_{1:k}):\sum_i \mathrm{E}X_i^2 \leq P} I(X_{1:k}; Y_{1:k}) \tag{904}$$

Canais em Paralelo IV

- Teremos o seguinte

$$\begin{aligned} I(X_{1:k};Y_{1:k}) &= h(Y_{1:k}) - h(Y_{1:k}|X_{1:k}) = h(Y_{1:k}) - h(Z_{1:k}|X_{1:k}) \\ &= h(Y_{1:k}) - h(Z_{1:k}) = h(Y_{1:k}) - \sum_{j=1}^{k} h(Z_j) \\ &\leq \sum_j (h(Y_j) - h(Z_j)) \leq \sum_j C_j = \sum_j \frac{1}{2}\log\left(1 + \frac{P_i}{N_i}\right) \end{aligned} \tag{905}$$

onde $P_i = \mathrm{E}X_i^2$ e $\sum_i P_i = P$.

- A igualdade ocorrerá quando

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} P_1 & 0 & \dots & & 0 \\ 0 & P_2 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \dots & & 0 & P_k \end{bmatrix}\right) \tag{906}$$

Canais em Paralelo V

- Teremos o seguinte problema de otimização para solucionar:

$$\begin{aligned} \underset{(P_1,P_2,\ldots,P_n)}{\text{maximizar}} \quad & \sum_j \frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right) \\ \text{sujeito a} \quad & \sum_i P_i = P \end{aligned} \tag{907}$$

- Lagrangiano será da forma

$$J(P_{1:n}) = \sum_j \frac{1}{2}\left(1+\frac{P_i}{N_i}\right) + \lambda\left(\sum_j P_j - P\right) \tag{908}$$

Otimização de Lagrange I

- Forma geral de uma problema de otimização:

$$\begin{aligned} \underset{x}{\text{minimizar}} \quad & f_0(x) \\ \text{sujeito a} \quad & f_i(x) \leq 0,\ i = 1, \ldots, m, \\ & h_i(x) = 0,\ i = 1, \ldots, p. \end{aligned} \tag{909}$$

- Forma do Lagrangiano

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{910}$$

e vamos definir

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) \tag{911}$$

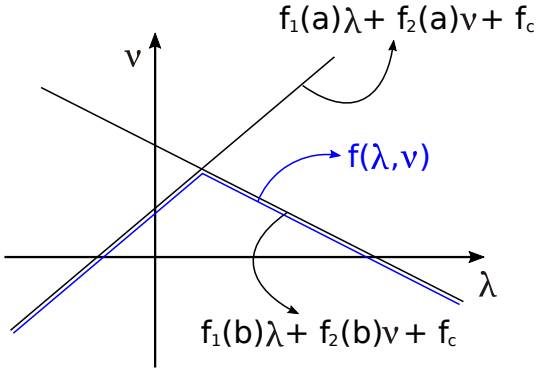
Otimização de Lagrange II

- g é côncava em (λ, ν), já que g é o ínfimo ponto-a-ponto sob uma transformação afim de (λ, ν), i.e.,

$$f(\lambda, \nu) = \min_{x \in \{a,b\}} (f_1(x)\lambda + f_2(x)\nu + f_c) \tag{912}$$

será côncava.

Otimização de Lagrange III



- Novamente, a definição de L é

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{913}$$

Otimização de Lagrange IV

- Considerando $\lambda \geq 0$, teremos, para o ponto ótimo x_{opt} (um ponto factível), que $\lambda_i f_i(x_{\text{opt}}) \leq 0$, como $\lambda_i \geq 0$, teremos $f_i(x_{\text{opt}}) \leq 0$, e $h_i(x_{\text{opt}}) = 0$.

$$\sum_{i=1}^{p} \nu_i h_i(x_{\text{opt}}) = 0 \qquad \sum_{i=1}^{m} \lambda_i f_i(x_{\text{opt}}) \leq 0 \tag{914}$$

- Definiremos $p_{\text{opt}} = L(x_{\text{opt}}, \lambda, \nu)$ e desta forma

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) \leq p_{\text{opt}} = L(x_{\text{opt}}, \lambda, \nu) \leq f(x_{\text{opt}}) \tag{915}$$

Exemplo de restrição severa I

- Considere a seguinte função

$$L^{\square}(x) = f_0(x) + \sum_{i=1}^{m} \left[\frac{1}{\mathbf{1}(f_i(x) \leq 0)} - 1 \right] + \sum_{i=1}^{p} \left[\frac{1}{\mathbf{1}(h_i(x) = 0)} - 1 \right] \tag{916}$$

- Note que, quando algum das restrições é violada, teremos $L^{\square} = \infty$. Se todas as restrições são satisfeitas, teremos $L^{\square}(x) = f_0(x)$.
- Para $\lambda_i > 0$ temos

$$\lambda_i f_i(x) \leq \left[\frac{1}{\mathbf{1}(f_i(x) \leq 0)} - 1 \right] \tag{917}$$

e para qualquer $\nu_i \in \mathbb{R}$, temos

$$\nu_i h_i(x) \leq \left[\frac{1}{\mathbf{1}(h_i(x) = 0)} - 1 \right] \tag{918}$$

Exemplo de restrição severa II

- Devemos então ter

$$L(x,\lambda,\nu) \leq L^{\square}(x) \quad \forall \lambda > 0, \nu \tag{919}$$

- Seja $x_{\text{opt}} = \min_x L^{\square}(x)$.
- Para quaisquer $\lambda > 0$ e ν,

$$\inf_x L(x,\lambda,\nu) \triangleq g(\lambda,\nu) \leq L(x_{\text{opt}},\lambda,\nu) \leq L^{\square}(x_{\text{opt}}) = f(x_{\text{opt}}) \tag{920}$$

- O melhor (maior) limite inferior, definirá valores ótimos duais, i.e.

$$(\lambda_{\text{opt}},\nu_{\text{opt}}) \in \underset{\lambda \geq 0,\nu}{\operatorname{argmax}}\, g(\lambda,\nu) \tag{921}$$

- Definiremos $d_{\text{opt}} = g(\lambda_{\text{opt}},\nu_{\text{opt}})$.

Dualidade fraca e intervalo de dualidade I

- A condição de dualidade fraca será dada quando o dual máximo não for maior do que o mínimo principal, i.e.,

$$d_{\text{opt}} \leq p_{\text{opt}} \Leftrightarrow \text{dualidade fraca} \tag{922}$$

- Note que a dualidade fraca será verdadeira pois, para quaisquer x', λ', ν' teremos

$$g(\lambda', \nu') = \inf_x L(x, \lambda', \nu') \leq L(x', \lambda', \nu') \leq \sup_{\lambda>0,\nu} L(x', \lambda, \nu) \leq L^{\square}(x') \tag{923}$$

- Assim,

$$d_{\text{opt}} = \sup_{\lambda>0,\nu} L(x, \lambda, \nu) \leq \inf_x \sup_{\lambda>0,\nu} L(x, \lambda, \nu) = \inf_x L^{\square}(x) = p_{\text{opt}} \tag{924}$$

- Teremos assim o seguinte intervalo de dualidade

$$p_{\text{opt}} - d_{\text{opt}} \geq 0 \tag{925}$$

Dualidade forte e intervalo de dualidade nulo I

- No caso de dualidade forte, teremos $p_{\text{opt}} = d_{\text{opt}}$.
- Podemos seguir maximizando o dual e minimizando o primário, se os dois se encontrarem, saberemos que achamos o ótimo (ou se estiverem próximos, teremos um limite para a qualidade da solução obtida).
- Infelizmente a dualidade forte nem sempre é satisfeita, enquanto a dualidade fraca é sempre verdadeira.
- Condições de Slater (condições para que a dualidade forte seja satisfeita): Se f_i são convexas e existem soluções que sejam estritamente factíveis (i.e., $\exists x$ tal que $\forall i$, $f_i(x) < 0$ e $A_i x = b_i$, de forma que $h_i(x) = A_i x - b_i$), então a dualidade forte será verdadeira.
- O intervalo de dualidade é importante para limitar a qualidade da solução

$$f_0(x) - p_{\text{opt}} \leq f_0(x) - g(\lambda, \nu) \quad \forall \lambda, \nu \tag{926}$$

- Note que $f_0(x) - g(\lambda, \nu)$ é o intervalo de dualidade associado entre x primário factível e o ponto dual factível (λ, ν).

Dualidade forte e intervalo de dualidade nulo II

- O intervalo de dualidade pode ser utilizado como critério para para o processo de otimização. Se o intervalo for nulo, então atingimos o ótimo.
- Assumindo dualidade forte, $p_{\text{opt}} = d_{\text{opt}}$, teremos

$$
\begin{aligned}
f_0(x_{\text{opt}}) &= g(\lambda_{\text{opt}}, \nu_{\text{opt}}) && (927)\\
&= \inf_x \left(f_0(x) + \sum_{i=1}^{n} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right) && (928)\\
&\leq f_0(x_{\text{opt}}) + \sum_{i=1}^{n} \lambda_i^* f_i(x_{\text{opt}}) + \sum_{i=1}^{p} \nu_i^* h_i(x_{\text{opt}}) && (929)\\
&\leq f_0(x_{\text{opt}}) && (930)
\end{aligned}
$$

onde $\lambda_{\text{opt}} = (\lambda_1^*, \lambda_2^*, \ldots)$, e de forma similar para ν_{opt}.

- 928 segue da definição
- 929 segue do ínfimo e dualidade fraca

Dualidade forte e intervalo de dualidade nulo III

- 930 segue de $\lambda_i^* \geq 0$, $f_i(x_{\text{opt}}) \leq 0$ e $h_i(x_{\text{opt}}) = 0$.
- teremos igualdade em 929 e 930 por causa da dualidade forte.

- Cada um dos termos do Lagrangiano será nulo, i.e.,

$$\left[\lambda_i^* f_i(x_{\text{opt}}) \leq 0, \forall i \quad \text{e} \quad \sum_{i=1}^{m} \lambda_i^* f_i(x_{\text{opt}}) = 0\right] \Rightarrow \lambda_i^* f_i(x_{\text{opt}}) = 0, \forall i \tag{931}$$

- Uma condição necessária para a otimalidade será

$$\text{se} \quad \lambda_i^* > 0 \quad \text{então} \quad f_i(x_{\text{opt}}) = 0 \tag{932}$$

ou

$$\text{se} \quad f_i(x_{\text{opt}}) < 0 \quad \text{então} \quad \lambda_i^* = 0 \tag{933}$$

uma das condições de Karush–Kuhn–Tucker (KKT) para otimalidade.

Dualidade forte e intervalo de dualidade nulo IV

- Cada um dos f_i será uma restrição ativa (i.e. $f_i(x_{\text{opt}}) = 0$) ou então serão não-ativos ($f_i(x_{\text{opt}}) \leq 0$) e assim $\lambda_i^* = 0$.
- Se todas as funções são diferenciáveis, então existe ∇f_0, ∇f_i e ∇h_i.
- Como $p_{\text{opt}} = \min_x \max_{\lambda>0,\nu} L(x,\lambda,\nu) = \min_x L(x,\lambda_{\text{opt}},\nu_{\text{opt}})$ teremos

$$\nabla_x L|_{x=x_{\text{opt}},\lambda=\lambda_{\text{opt}},\nu=\nu_{\text{opt}}} = 0 \tag{934}$$

- Então, uma outra condição para otimalidade é

$$\nabla_x f_0(x_{\text{opt}}) + \sum_{i=1}^{m} \lambda_i^* \nabla_x f_i(x_{\text{opt}}) + \sum_{i=1}^{p} \nu_i^* \nabla_x h_i(x_{\text{opt}}) = 0 \tag{935}$$

Dualidade forte e intervalo de dualidade nulo V

- As condições de KKT para otimalidade serão

$$f_i(x_{\text{opt}}) \leq 0 \quad \text{para} \quad i = 1, \ldots, m \tag{936}$$

$$h_i(x_{\text{opt}}) = 0 \quad \text{para} \quad i = 1, \ldots, p \tag{937}$$

$$\lambda_i^* \geq 0 \quad \text{para} \quad i = 1, \ldots, m \tag{938}$$

$$\lambda_i^* f_i(x_{\text{opt}}) = 0 \quad \text{para} \quad i = 1, \ldots, m \tag{939}$$

e

$$\nabla_x L|_{x=x_{\text{opt}}, \lambda=\lambda_{\text{opt}}, \nu=\nu_{\text{opt}}} = 0 \tag{940}$$

aonde utilizamos a notação $\lambda_{\text{opt}} = \lambda^*$ e $\nu_{\text{opt}} = \nu^*$.

Canais Paralelo I

- Queremos solucionar o problema

$$\begin{aligned} &\underset{(P_1,P_2,\ldots,P_n)}{\text{maximizar}} && \sum_j \frac{1}{2}\log\left(1+\frac{P_i}{N_i}\right) \\ &\text{sujeito a} && \sum_i P_i = P \end{aligned} \tag{941}$$

- ou na forma do Lagrangiano

$$J(P_{1:n}) = \sum_j \frac{1}{2}\left(1+\frac{P_i}{N_i}\right) + \lambda\left(\sum_j P_j - P\right) \tag{942}$$

- $x = (P_1, P_2, \ldots, P_m)$ é um vetor de potências
- N_i ruído dado em cada canal (σ_i^2)

Canais Paralelo II

- Queremos

$$\text{minimizar} \quad -\sum_{i=1}^{k} \log(1 + P_i/N_i) \tag{943}$$

  sujeito às desigualdades de restrição $P_i \geq 0$ (então $f_i(P_i) = -P_i$) e igualdades de restrição $\sum_{i=1}^{k} P_i = P$ (i.e. $h = (\sum_{j=1}^{n} P_i - P) = 0$).
- Este problema é convexo.
- Então existe um ponto estritamente factível. Desta forma a dualidade forte é válida, assim como as condições de KKT para a otimizalidade.

Canais Paralelo III

- Teremos o seguinte Lagrangeano:

$$L(x,\lambda,\nu) = -\sum_{i=1}^{k}\log(1+P_i/N_i) - \sum_{i=1}^{k}\lambda_i P_i + \nu\left(\sum_{i=1}^{k}P_i - P\right) \tag{944}$$

- As codiçoes de KKT serao:

$$\forall i : P_i^* \geq 0, \quad \sum_i P_i^* = P, \quad \lambda_i^* \geq 0 \forall i, \quad \lambda_i^* P_i^* = 0 \tag{945}$$

e também, $\forall i$

$$-\frac{1}{(1+P_i/N_i)}\frac{1}{N_i} - \lambda_i^* + \nu^* = 0 \tag{946}$$

Canais Paralelo IV

- A partir das condicoes do gradiente do Lagrangeano, podemos ainda obter

$$-\frac{1}{N_i + P_i} - \lambda_i^* + \nu^* = 0 \tag{947}$$

$$-\frac{1}{N_i + P_i} + \nu^* = \lambda_i^* \geq 0 \tag{948}$$

- Podemos eliminar λ_i^* para obter as condiçoes de KKT na forma

$$\forall i : P_i^* \geq 0 \quad \sum_i P_i^* = P \tag{949}$$

$$\left(\nu^* - \frac{1}{N_i + P_i}\right) P_i^* = 0 \quad \nu^* \geq \frac{1}{N_i + P_i^*} \tag{950}$$

Canais Paralelo V

Temos entao dois casos:

Caso 1 :

- a partir da condiçao $\nu^* \geq \frac{1}{N_i+P_i^*}$, se $\nu^* < 1/N_i$, entao devemos ter $P_i^* > 0$ para alcançar esta condiçao.
- Em tal caso, como temos $\left(\nu^* - \frac{1}{N_i+P_i}\right) P_i^* = 0$, devemos ter $\nu^* = \frac{1}{N_i+P_i^*}$.
- Assim, $P_i^* = \frac{1}{\nu^*} - N_i$.

Caso 2 :

- se $\nu^* \geq 1/N_i$, entao $P_i^* = 0$, já que no caso contrário ($P_i^* > 0$) significaria

$$\underbrace{\left(\nu^* - \frac{1}{N_i + P_i}\right)}_{>0 \text{ já que } \nu^* \geq 1/N_i \text{ e } P_i^* > 0} \times \underbrace{P_i^*}_{>0} > 0 \tag{951}$$

Canais Paralelo VI

- P_i^* deverá ser da forma

$$P_i^* = \begin{cases} 1/\nu^* - N_i & \text{se } \nu^* < 1/N_i, \\ 0 & \text{se } \nu^* \geq 1/N_i. \end{cases} \tag{952}$$

$$= \max\{0, 1/\nu^* - N_i\} \tag{953}$$

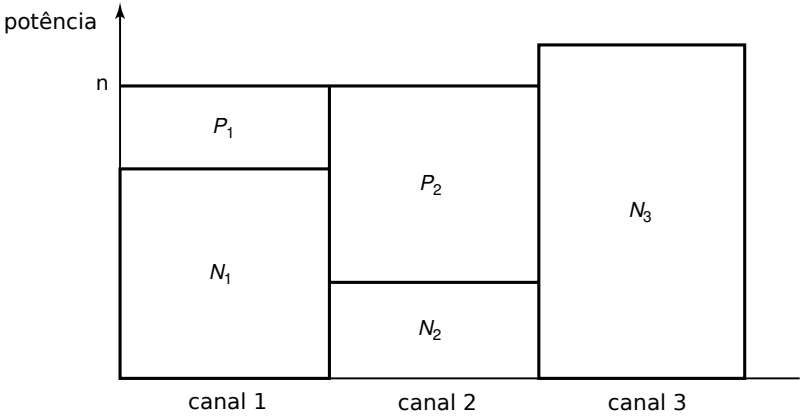
- com a ultima restriçao teremos

$$\sum_i \left(\frac{1}{\nu^*} - N_i\right)^+ = P \tag{954}$$

onde $a^+ = \max\{0, a\}$.

- isto leva à ideia da água enchendo os canais em paralelo.

Canais Paralelo VII



Canais Paralelo VIII

A capacidade final será

$$C_n = \frac{1}{2}\sum_{j=1}^{k}\log(1+P_i/N_i) \quad (955)$$

$$= \frac{1}{2}\sum_{j=1}^{k}\log\left(1+\frac{(1/\nu^*-N_i)^+}{N_i}\right) \quad (956)$$

Em unidades de bits por transmissao (bits por transmissao em canal simples, tomando a média):

$$C_n = \frac{1}{2n}\sum_{j=1}^{k}\log\left(1+\frac{(1/\nu^*-N_i)^+}{N_i}\right) \text{ bits por transmissao} \quad (957)$$

Bilmes, J. A. (2013). Lecture notes on information theory.

Hartley, R. V. L. (1928). Transmission of information1. *Bell System Technical Journal*, 7(3):535–563.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK ; New York.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

Wikipedia (2020a). Etaoin shrdlu. https://en.wikipedia.org/wiki/Etaoin_shrdlu. [Online; accessed 10-August-2020].

Wikipedia (2020b). Jean-dominique bauby. https://en.wikipedia.org/wiki/Jean-Dominique_Bauby. [Online; accessed 10-August-2020].

Wikipedia (2020c). Letter frequency. https://en.wikipedia.org/wiki/Letter_frequency. [Online; accessed 10-August-2020].

Wikipedia (2020d). Morse code. https://en.wikipedia.org/wiki/Morse_code. [Online; accessed 10-August-2020].