Elementos de Teoria da Informação

A teoria da informação nasceu oficialmente em 1948, com o célebre artigo de Claude Shannon: "Mathematical Theory of Communication". A teoria foi desenvolvida em conexão com o problema da transmissão da informação em canais de comunicação. A principal preocupação da teoria era procurar por um método de transmitir informação através de um canal com ruído com eficiência e confiabilidade, ou seja, a uma taxa alta de transmissão com o mínimo de erros.

Informação como probabilidade é um conceito muito geral, qualitativo, impreciso e muito subjetivo. A mesma "informação" pode ter significados distintos, efeitos e valores distintos para pessoas diferentes. Mesmo assim, tal como a teoria das probabilidades que se desenvolveu a partir de conceitos imprecisos e subjetivos, a teoria da informação se tornou uma teoria quantitativa, precisa, objetiva e muito útil.

A teoria da informação é usada para se estimar a "melhor" distribuição de probabilidade,

bem como para interpretar os conceitos fundamentais na teoria da mecânica estatística da matéria. A significado da palavra informação neste caso não estar relacionado à informação em si mesma, nem à quantidade de informação transportada pela mensagem, mas sim ao tamanho da mensagem que transporta informação.

Introdução Qualitativa à Teoria da Informação

Um jogo bastante popular é o jogo das 20 perguntas, no qual um jogador escolhe uma pessoa e o outro jogador deve descobrir qual é essa pessoa através de questões que só admitem respostas binárias: "sim" ou "não" ("0" ou "1"). O jogador pode fazer no máximo 20 questões. Suponha que um jogador tenha escolhido Einstein.

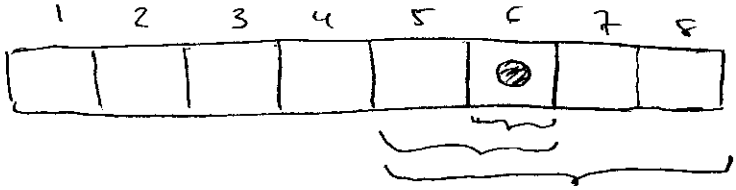
Existem dois tipos de estratégias para se fazer as perguntas:

Estratégia Tola	Estratégia Inteligente
1) É o Obama?	1) É um homem
2) É o Lula?	2) Ele está vivo?
3) Sou eu?	3) Ele é um político?
4) Ela é a Marilyn Monroe?	4) Ele é um cientista?
5) É você?	5) Ele é famoso?
6) Ele é Mozart?	6) Ele é Einstein?
7) Ele é Bohr?	

Um jogador de muita sorte pode ganhar o jogo na primeira pergunta utilizando a estratégia tola e um jogador nunca ganhará de primeira utilizando a estratégia inteligente. Na média as chances de um jogador ganhar usando a estratégia inteligente são muito maiores que utilizando a estratégia tola. A razão da estratégia inteligente ser muito mais eficiente é que a cada pergunta um número grande de possibilidades incorretas são eliminadas, ao passo que na estratégia tola apenas uma possibilidade é eliminada por pergunta. Adquirimos muito mais informação por pergunta na estratégia inteligente do que na estratégia tola.

Vamos analisar agora um jogo muito mais simples e mais passível de um tratamento objetivo e quantitativo.

Consideremos, por exemplo, 8 caixas iguais. Em uma das caixas há uma moeda escondida. O objetivo do jogo é descobrir onde se encontra a moeda através de perguntas binárias.



Novamente neste caso temos uma estratégia tola e uma inteligente.

Estratégia Tola

1) Está na 1ª caixa?
2) Está na 2ª caixa?
3) Está na 3ª caixa?
4) Está na 4 caixa?

Estratégia Inteligente

1) Está na metade da direita de 8 caixas?
2) Está na metade da direita das 4 caixas restantes?
3) Está na metade da direita das 2 caixas restantes?
4) Eu sei a resposta?

Também neste caso, o jogador que utiliza a estratégia tola pode acertar a localização da moeda na primeira pergunta, mas na média esta estratégia requer um número bem maior que o da estratégia inteligente, que neste caso necessitou de apenas 3 perguntas.

Podemos verificar facilmente que o número de perguntas necessárias utilizando-se a estratégia inteligente no caso de 16 caixas é 4 e para 32 caixas são necessárias 5 perguntas.

Temos então que o número de questões pode ser expresso pela equação:

$$H(n) = \log_2 n = \log_2 2^m = m$$

onde n é o número de caixas e m o número de perguntas binárias. Em outras palavras, precisamos de uma mensagem de m bits para nos informar onde a moeda se encontra. No caso do exemplo das 8 caixas, as respostas às 3 questões seriam: sim, não, sim ou 1 0 1.

Como todas as caixas são iguais, a probabilidade da moeda estar na i-ésima caixa é:

$$p_i = \frac{1}{n}$$

então temos,

$$H(n) = \frac{n}{n} \log_2 n = -n \frac{1}{n} \log_2 \frac{1}{n}$$

ou

$$H(p_1 \cdots p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Esta é a famosa expressão obtida por Shannon para quantificar a incerteza ou entropia associada

Shannon empregou o logaritmo na base 2 na sua expressão para que a expressão fornecesse a incerteza ou entropia em bits ou dígitos binários. Podemos mudar a base dos logaritmos facilmente, simplesmente multiplicando a expressão por uma constante. Por exemplo,

$$H(p_i) = -\sum_{i=1}^{n} p_i \log_2 p_i = -\log_2 e \sum_{i=1}^{n} p_i \ln p_i$$

Portanto, a menos de uma constante multiplicativa positiva, a entropia da informação é dada por:

$$H(p_i) = -K \sum_{i=1}^{n} p_i \log p_i$$

na qual a base dos logaritmos não necessita estar especificada.

Na verdade outras bases podem ser utilizadas na definição da entropia da informação. O que muda é a unidade em que ela é expressa. Por exemplo, se usarmos a base 10 a entropia será dada na unidade ban ao invés de bits. No caso da base neperiana e, a unidade tem vários nomes: nat, nit ou nepit. Apesar das unidades serem diferentes, as propriedades fundamentais são exatamente as mesmas, independentemente da base escolhida.

Propriedades da Entropia de Informação

Na teoria matemática da informação desenvolvida por Shannon, inicia-se por considerar uma variável aleatória X, ou um experimento (ou um jogo). A distribuição de probabilidade de X, $p_1, p_2, \cdots p_n$ é assumida como sendo conhecida. A questão colocada por Shannon é a seguinte:

"Can we find a measure of how much 'choice' is involved in the selection of the event, or how much uncertain we are of the outcome?"

Shannon então considerou que tal função, $H(p_1, p_2 \cdots p_n)$, existe e é razoável que ela tenha as seguintes propriedades:

i) H deve ser contínua em todos os p_i.

ii) Se todos os p_i são iguais, ou seja, $p_i = 1/n$, então o valor de H deve ser máximo e este valor máximo deve ser monotonicamente crescente com n.

iii) Se uma escolha for quebrada em escolhas sucessivas, a quantidade H deve ser uma soma ponderada dos valores individuais de H.

Estas não são apenas propriedades desejáveis de H, mas também são propriedades razoáveis que se poderia esperar de tal quantidade.

A primeira propriedade é razoável no sentido que se fizermos variações arbitrariamente pequenas nas probabilidades, então esperamos que a variação na incerteza seja pequena.

A segunda propriedade também é plausível. Para um dado n fixo, se a distribuição for uniforme, nós temos a informação mínima sobre o resultado do experimento, ou seja, a máxima incerteza sobre esse resultado. Claramente, quanto maior for o número n, maior será a informação necessária.

A terceira propriedade é algumas vezes referida como a da independência do agrupamento dos eventos. Este requerimento é equivalente a se dizer a informação ausente deve depender apenas da distribuição $p_1, p_2, \cdots p_n$ e não da maneira específica pela qual a informação é adquirida, por exemplo, perguntando-se questões binárias usando diferente estratégias.

Shannon provou que a única função que satisfaz esses 3 requisitos é:

$$H(p_1, p_2, \cdots p_n) = -K \sum_{i=1}^{n} p_i \log p_i$$

onde K é uma constante positiva qualquer e independente da base do logaritmo. Muitas vezes fazemos $K = 1$.

O caso mais simples de dois resultados

O caso mais simples ainda em que $\Omega = A$ e $P(A) = 1$, o resultado é único e certo, neste caso $H = 0$, não há incerteza.

O caso de dois resultados possíveis temos A_1 e $A_2 = \bar{A}_1$ com probabilidades p_1 e p_2.

Neste caso, podemos escrever: $p_1 = p$ e $p_2 = 1-p$.

Então:

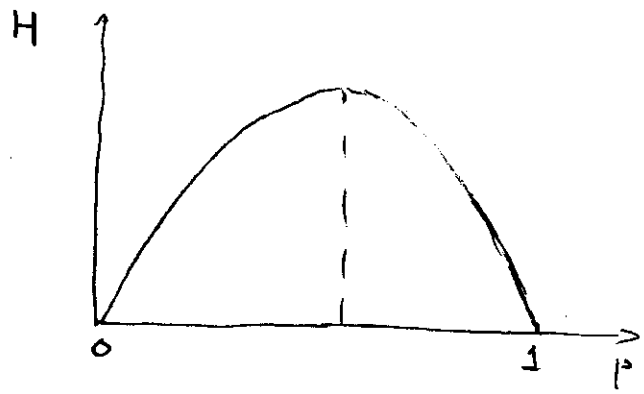
$$H = -\sum_i p_i \log_2 p_i = -p \log_2 p - (1-p)\log_2(1-p)$$

Se $p \to 0$, então $H \to 0$.

Lembremos que: $\lim_{x\to 0}[x \ln x] = \lim_{x\to 0} \frac{\ln x}{1/x} =$

$$= \lim_{x\to 0} \frac{\frac{d}{dx}[\ln x]}{\frac{d}{dx}\left[\frac{1}{x}\right]} = \lim_{x\to 0} \frac{1/x}{-1/x^2} = 0$$

O mesmo vale para $p \to 1$, $H \to 0$.



Se $p = \frac{1}{2}$, no caso da base 2, temos

$$H = -\frac{1}{2}\log_2 \frac{1}{2} - \left(1-\frac{1}{2}\right)\log_2\left(1-\frac{1}{2}\right)$$

$$= \frac{1}{2} + \frac{1}{2} = 1 \text{ bit}$$

Temos 1 bit de informação ausente.

Propriedades de H para o caso geral de n resultados

Consideremos a variável aleatória X com a distribuição de probabilidade dada por $P_X(i) = P\{X(\omega) = i\} = p_i$. O subescrito X é normalmente omitido quando sabemos qual a variável aleatória está sendo considerada. Então escrevemos:

$$H = -\sum_{i=1}^{n} p_i \log p_i$$

com a condição de normalização

$$\sum_{i=1}^{n} p_i = 1$$

A propriedade mais importante de H é que ela é máxima quando todos os p_i's são iguais. Podemos provar isso usando o método dos multiplicadores de Lagrange, para determinarmos o máximo de uma função sujeita a vínculos.

Seja a função auxiliar:

$$F = H(p_1, \ldots, p_n) + \lambda\left[\sum_{i} p_i - 1\right]$$

Calculando as derivadas parciais de F com respeito a cada um dos p_i temos:

$$\frac{\partial F}{\partial p_i} = -\log p_i - 1 + \lambda = 0$$

ou

$$p_i = \exp(\lambda - 1)$$

substituindo esta última equação na condição de normalização obtemos:

$$1 = \sum_{i=1}^{n} p_i = \exp(\lambda - 1) \sum_{i=1}^{n} 1 = n \exp(\lambda - 1)$$

Portanto, temos

$$\exp(\lambda - 1) = \frac{1}{n}$$

ou

$$p_i = \frac{1}{n}$$

Este resultado é muito importante, o qual nos diz que o valor máximo de H, sujeita apenas à condição de normalização dos p_i's, é obtido quando a distribuição é uniforme. Esta é uma generalização do caso de duas variáveis aleatórias.

O valor máximo que H assume é:

$$H_{max} = -\sum_{i=1}^{n} p_i \log p_i = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$$

É claro que, quando há n resultados igualmente prováveis, a quantidade de informação ausente é tanto maior quanto maior for o número de resultados

No caso em que a distribuição não é uniforme, é claro que a informação ausente será menor que o valor máximo de H. Isto significa que em média um número menor de questões terá que ser perguntado para que a informação ausente seja obtida. Um caso limite de distribuição não uniforme é aquele em que a quantidade H é nula se e somente se a ocorrência ou não ocorrência de um evento é certa. Por exemplo, $p_1 = 1$ e $p_i = 0$, para todos $i = 2, \ldots, n$. Isto é claro, porque como temos certeza da ocorrência de um dado evento e, portanto, da não ocorrência de outros eventos, não há informação ausente.

Consideremos agora que temos duas variáveis aleatórias, X e Y com distribuições $P_X(i) = P\{X = x_i\}$ e $P_Y(j) = P\{Y = y_j\}$, $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, m$. Seja $P(i,j)$ a probabilidade conjunta de ocorrência dos eventos $\{X = x_i\}$ e $\{Y = y_j\}$. A função H para a distribuição conjunta $P(i,j)$ é:

$$H(X,Y) = -\sum_{i,j}^{'} P(i,j) \log P(i,j)$$

As probabilidades marginais serão:

$$p_i = \sum_{j=1}^{m} P(i,j) = P_x(i)$$

e

$$q_j = \sum_{i=1}^{n} P(i,j) = P_y(j)$$

A informação associada com as variáveis aleatórias X e Y são:

$$H(X) = -\sum_{i=1}^{n} P_x(i) \log P_x(i) = -\sum_{ij} P(i,j) \log \sum_{j=1}^{m} P(i,j)$$

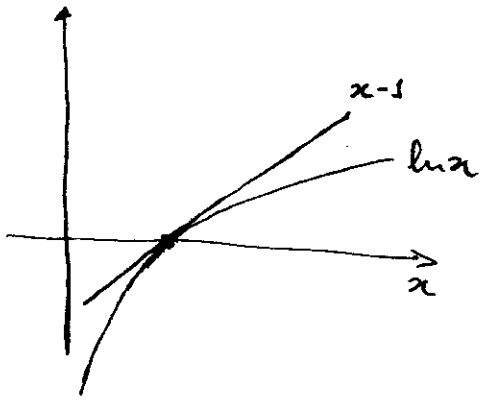
e

$$H(Y) = -\sum_{j=1}^{m} P_y(j) \log P_y(j) = -\sum_{ij} P(i,j) \log \sum_{i=1}^{n} P(i,j)$$

Se duas distribuições quaisquer $\{p_i\}$ e $\{q_j\}$ forem normalizadas: $\sum_{i=1}^{n} p_i = 1$ e $\sum_{i=1}^{n} q_i = 1$, a seguinte desigualdade é válida:

$$H(q_1, \cdots q_n) = -\sum_{i=1}^{n} q_i \log q_i \leq -\sum_{i=1}^{n} q_i \log p_i$$

Este resultado pode ser demonstrado através da desigualdade $\ln x \leq x - 1$. Esta desigualdade é verdadeira para qualquer $x > 0$. Isto pode ser visto pelo fato que $\frac{d}{dx}(\ln x)_{x=1} = \frac{1}{x}\Big|_{x=1} = 1$ e, portanto, a reta $x-1$ é tangente a curva $\ln x$ em $x = 1$ e pela concavidade de $\ln x$, todos os valores de $\ln x$ serão menores que $x - 1$.



Retornando ao nosso problema, escolhendo $x = \frac{p_i}{q_i}$, nós teremos:

$$\ln \frac{p_i}{q_i} \leq \frac{p_i}{q_i} - 1$$

Multiplicando a desigualdade por q_i e somando sobre i, nós obtemos,

$$\sum_{i=1}^{n} q_i \ln \frac{p_i}{q_i} \leq \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} q_i = 0$$

ou

$$\sum_{i=1}^{n} q_i \ln p_i - \sum_{i=1}^{n} q_i \ln q_i \leq 0$$

ou

$$-\sum_{i=1}^{n} q_i \ln q_i \leq -\sum_{i=1}^{n} q_i \ln p_i$$

Esta última expressão pode ser transformada para logaritmos de qualquer base:

$$-\sum_{i=1}^{n} q_i \log q_i \leq -\sum_{i=1}^{n} q_i \log p_i$$

Obs: Esta desigualdade é válida para o caso em que as distribuições se referem à duas variáveis aleatórias

Pelo que já foi visto:

$$H(X) + H(Y) =$$

$$= -\sum_{i}{}' P_x(i) \log P_x(i) - \sum_{j}{}' P_y(j) \log P_y(j)$$

$$= -\sum_{i,j}{}' P(i,j) \log P_x(i) - \sum_{i,j}{}' P(i,j) \log P_y(j)$$

$$= -\sum_{i,j}{}' P(i,j) \left[\log P_x(i) + \log P_y(j) \right]$$

$$= -\sum_{i,j}{}' P(i,j) \log \left[P_x(i) P_y(j) \right]$$

Mas pela desigualdade vista anteriormente

$$-\sum_{i,j}{}' P(i,j) \log \left[P_x(i) P_y(j) \right] \geq -\sum_{i,j}{}' P(i,j) \log P(i,j)$$

Portanto,

$$H(x) + H(y) \geq H(x,y)$$

Quando as duas variáveis aleatórias são independentes temos:

$$P(i,j) = P_x(i) P_y(j)$$

Desta forma:

$$-\sum_{i,j} P(i,j) \log_2 \left[P_X(i) P_Y(j) \right] = \sum_{i,j} P(i,j) \log_2 P(i,j)$$

e neste caso vale a igualdade

$$H(X) + H(Y) = H(X,Y)$$

Estes dois últimos resultados simplesmente significam que se tivermos dois experimentos, cujos resultados são independentes, então a informação ausente sobre o resultado dos dois experimentos é a soma das informações ausentes sobre os resultados de cada um dos experimentos. Por outro lado, se há dependência entre os dois conjuntos de resultados, então a informação sobre o experimento conjunto (X,Y) é sempre menor que a informação ausente dos dois experimentos separadamente.

Para experimentos em que há dependência, nós utilizamos probabilidades condicionais.

$$P(y_j | x_i) = \frac{P(x_i \cdot y_j)}{P(x_i)}$$

para definir a quantidade condicional correspondente:

$$H(Y|X) = \sum_i' P(x_i)\, H(Y|x_i)$$

$$= -\sum_i' P(x_i) \sum_j' P(y_j|x_i) \log P(y_j|x_i)$$

$$= -\sum_{i,j}' P(x_i \cdot y_j) \log P(y_j|x_i)$$

$$= -\sum_{i,j}' P(x_i \cdot y_j) \log P(x_i \cdot y_j) +$$

$$+ \sum_{i,j}' P(x_i \cdot y_j) \log P(x_i)$$

$$= H(X,Y) - H(X)$$

Portanto, $H(Y|X)$ mede a diferença entre a informação ausente sobre X e Y e a informação ausente sobre X.

Isto pode ser reescrito sob a forma:

$$H(X,Y) = H(X) + H(Y|X)$$

$$= H(Y) + H(X|Y)$$

o que significa que a incerteza em dois experimentos é a soma da incerteza em um desses experimentos com a incerteza quando o resultado do outro experimento é conhecida.

Das relações:

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

e

$$H(X,Y) = H(X) + H(Y|X)$$

temos

$$H(X) + H(Y) \geq H(X) + H(Y|X)$$

e, portanto,

$$H(Y|X) \leq H(Y)$$

que significa que a informação ausente sobre Y nunca pode crescer através do conhecimento sobre a outra variável X. Alternativamente, $H(Y|X)$ é a incerteza média que resta sobre Y quando X é conhecido.

Propriedade de Consistência da Entropia de Informação

O terceiro requerimento que a entropia de informação deve satisfazer é a consistência. Esta condição essencialmente diz que a quantidade de informação em uma dada distribuição $(p_1, p_2, \cdots p_n)$ é independente do caminho, ou do número de passos nós escolhemos dar para obter esta informação. Em outras palavras, a quantidade de informação é obtida independente da maneira ou número de passos usados para adquirir a informação. Na sua forma mais geral o enunciado é formulado da seguinte forma.

Suponhamos que temos n resultados $A_1, \cdots A_n$ de um dado experimento, cujas probabilidades correspondentes são $p_1, p_2, \cdots p_n$. Podemos reagrupar os resultados da seguinte maneira.

$$\{A_1, A_2, A_3, A_4, A_5, A_6, A_7, \cdots, A_n\}$$

$$\{A_1, A_2, A_3\}, \{A_4, A_5, A_6, A_7\}, \cdots, \{A_{n-2}, A_{n-1}, A_n\}$$

$$A_1', A_2', \cdots A_r'$$

A_1' é o novo evento que consiste dos eventos originais A_1, A_2, A_3. As probabilidades correspondentes são $P_1', P_2', \cdots P_n'$.

Então, do conjunto inicial de n resultados, nós construímos o novo conjunto de r eventos $\{A_1', A_2', \cdots A_r'\}$. Consideremos que todos os A_i sejam mutuamente exclusivos, então as novas probabilidades são dadas por:

$$P_1' = \sum_{i=1}^{m_1}{}' p_i \;, \quad P_2' = \sum_{i=m_1+1}^{m_1+m_2}{}' p_i \;, \cdots$$

Então, o evento A_1' consiste de m_1 dos eventos originais com probabilidade P_1', A_2' consiste de m_2 eventos originais com probabilidade P_2', e assim por diante. Ao todo, nós repartimos os n eventos em r grupos, cada um contendo m_k $(k=1, \cdots r)$ dos eventos originais, de forma que

$$\sum_{k=1}^{r}{}' m_k = n$$

A condição de consistência é escrita como:

$$H(p_1, p_2, \dots p_n) = H(p_1', \dots p_n') + \sum_{k=1}^{n} p_k' H\left(\frac{p_1^k}{p_k'}, \dots \frac{p_{m_k}^k}{p_k'}\right)$$

onde p_i^k é i-ésima probabilidade do k-ésimo grupo, referentes aos eventos originais.

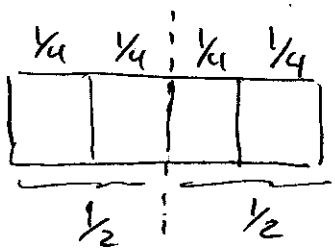
O significado dessa equação é o seguinte. A entropia de informação do sistema original $H(p_1, \dots, p_n)$ é igual a entropia do novo conjunto de n eventos mais a entropia de informação média associada a cada um dos grupos.

Vejamos um exemplo. Consideremos o caso das quatro caixas, sendo que em uma delas está escondida uma moeda. Nós consideramos que as probabilidades de encontrarmos a moeda em qualquer uma das caixas são iguais a $1/4$. A informação ausente neste caso é:

$$H\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) = -\sum_{i=1}^{4} \frac{1}{4} \log_2 4 = \log_2 4 = 2$$

Ou seja, nós precisamos de 2 bits de informação para localizar a moeda. Nós podemos obter essa informação por diferentes caminhos. A condição de consistência significa que essa quantidade de informação deve ser a mesma independente da estratégia usada para obtê-la.

A primeira nota é dividir o número total de caixas em duas metades, cada uma tendo probabilidade $\frac{1}{2}$.



Neste caso, a condição de consistência fica:

$$H\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right) = H\left(\frac{1}{2},\frac{1}{2}\right) + \left[\frac{1}{2}H\left(\frac{1/4}{1/2},\frac{1/4}{1/2}\right) + \frac{1}{2}H\left(\frac{1/4}{1/2},\frac{1/4}{1/2}\right)\right]$$

$$= H\left(\frac{1}{2},\frac{1}{2}\right) + \left[\frac{1}{2}H\left(\frac{1}{2},\frac{1}{2}\right) + \frac{1}{2}H\left(\frac{1}{2},\frac{1}{2}\right)\right]$$

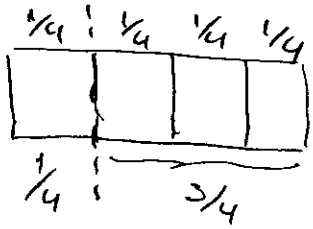
$$= 2\,H\left(\frac{1}{2},\frac{1}{2}\right)$$

mas $H\left(\frac{1}{2},\frac{1}{2}\right) = 1$ bit

Portanto,

$$H\left(\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}\right) = 2 \text{ bits.}$$

Suponhamos agora uma nota diferente. Ao invés de dividir em duas metades, nós dividimos em dois grupos, 1 caixa e 3 caixas. Neste caso temos,



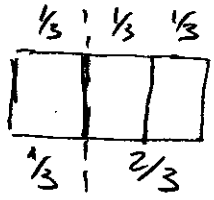
$$H\left(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4}\right) = H\left(\tfrac{1}{4},\tfrac{3}{4}\right) + \left[\tfrac{1}{4} H\left(\tfrac{1/4}{1/4}\right) + \tfrac{3}{4} H\left(\tfrac{1/4}{3/4},\tfrac{1/4}{3/4},\tfrac{1/4}{3/4}\right)\right]$$

$$= H\left(\tfrac{1}{4},\tfrac{3}{4}\right) + \left[\tfrac{1}{4} H(1) + \tfrac{3}{4} H\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)\right]$$

mas $H(1) = -1 \ln 1 = 0$, portanto,

$$H\left(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4}\right) = H\left(\tfrac{1}{4},\tfrac{3}{4}\right) + \tfrac{3}{4} H\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right)$$

Podemos agora subdividir o grupo de 3 caixas em dois grupos novamente, 1 caixa e 2 caixas.



Então,

$$H\left(\tfrac{1}{3},\tfrac{1}{3},\tfrac{1}{3}\right) = H\left(\tfrac{1}{3},\tfrac{2}{3}\right) + \left[\tfrac{1}{3} H\left(\tfrac{1/3}{1/3}\right) + \tfrac{2}{3} H\left(\tfrac{1/3}{2/3},\tfrac{1/3}{2/3}\right)\right]$$

$$= H\left(\tfrac{1}{3},\tfrac{2}{3}\right) + \left[\tfrac{1}{3} H(1) + \tfrac{2}{3} H\left(\tfrac{1}{2},\tfrac{1}{2}\right)\right]$$

Juntando os termos vamos obter,

$$H\left(\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4},\tfrac{1}{4}\right) = H\left(\tfrac{1}{4},\tfrac{3}{4}\right) + \tfrac{3}{4}\left[H\left(\tfrac{1}{3},\tfrac{2}{3}\right) + \left[\tfrac{1}{3} \overset{0}{H(1)} + \tfrac{2}{3} H\left(\tfrac{1}{2},\tfrac{1}{2}\right)\right]\right]$$

$$= \underbrace{H\left(\tfrac{1}{4},\tfrac{3}{4}\right)}_{0.8113} + \underbrace{\tfrac{3}{4} H\left(\tfrac{1}{3},\tfrac{2}{3}\right)}_{0.6887} + \underbrace{\tfrac{1}{2} H\left(\tfrac{1}{2},\tfrac{1}{2}\right)}_{\frac{1}{2}}$$

$$= 2 \text{ bits}$$

A variação não inteira da entropia neste caso evidencia a assimetria do caminho seguido. Diferentemente do caso simétrico, neste caso a moeda pode ser encontrada [illegible] nenhuma. Significando que em média na

O Caso de Distribuições Contínuas de Probabilidade

No caso de termos que a distribuição de probabilidade da variável aleatória é contínua:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx$$

a entropia de informação é definida por:

$$H = -\int_{-\infty}^{\infty} f(x) \log f(x)\,dx$$

Esta definição tem problemas, mesmo no caso em que os limites de integração sejam finitos.
Consideremos o caso em que a variável aleatória X possa assumir qualquer valor no intervalo (a,b) e que exista uma densidade de probabilidade $f(x)$ tal que

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx$$

e

$$\int_a^b f(x)\,dx = 1$$

Vamos dividir o intervalo (a,b) em n intervalos, cada um de comprimento:

$$\delta = \frac{(b-a)}{n}$$

de forma que,

$$x_1 = a \quad , \quad x_i = a + (i-1)\delta \quad , \quad x_{n+1} = a + n\delta = b$$

sendo que $i = 1, 2, \cdots, n+1$

Então, a probabilidade,

$$P(i,n) = \int_{x_i}^{x_{i+1}} f(x)\, dx$$

é a probabilidade de X estar entre x_i e x_{i+1}, para uma subdivisão do intervalo (a,b) em n intervalos.

A informação ausente associada a $P(i,n)$ é dada por

$$H(n) = -\sum_{i=1}^{n} P(i,n) \log P(i,n)$$

Como neste caso, $H(n)$ é definida para um valor finito de n, a expressão acima não traz em si nenhum problema.

Vamos agora substituir $P(i,n)$, na expressão de $H(n)$, por sua definição em termos da integral de $f(x)$:

$$H(n) = -\sum_{i=1}^{n}\left[\int_{x_i}^{x_{i+1}} f(x)\,dx\right]\log\left[\int_{x_i}^{x_{i+1}} f(x)\,dx\right]$$

$$= -\sum_{i=1}^{n}\left[\bar{f}(i,n)\,\delta\right]\log\left[\bar{f}(i,n)\,\delta\right]$$

$$= -\sum_{i=1}^{n}\left[\bar{f}(i,n)\,\frac{(b-a)}{n}\right]\log\left[\bar{f}(i,n)\right]$$

$$-\sum_{i=1}^{n}\left[\bar{f}(i,n)\,\frac{(b-a)}{n}\right]\log\left[\frac{b-a}{n}\right]$$

onde $\bar{f}(i,n)$ é algum valor da função $f(x)$ entre $f(x_i)$ e $f(x_{i+1})$, para um dado valor de n.

Quando $n \to \infty$ temos,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \bar{f}(i,n)\,\frac{(b-a)}{n} = \int_a^b f(x)\,dx = 1$$

$$\lim_{n\to\infty} \sum_{i=1}^{n} \bar{f}(i,n)\log\left[\bar{f}(i,n)\right]\frac{(b-a)}{n} = \int_a^b f(x)\log f(x)\,dx$$

Os dois limites acima são essencialmente a definição da integral de Riemann. Devemos notar que a integral $\int_a^b f(x)\log f(x)\,dx$ pode ser positiva ou negativa.

Temos então que,

$$H = \lim_{n\to\infty} H(n) = -\int_a^b f(x)\log f(x)\,dx - \lim_{n\to\infty} \log\left[\frac{b-a}{n}\right]$$

O segundo termo da expressão acima claramente diverge quando $n\to\infty$. A razão da divergência é que quanto maior for o valor de n, maior o número de intervalos e maior será a informação necessária para se localizar um ponto no intervalo (a,b). Notemos, entretanto, que o termo divergente não depende da distribuição $f(x)$. Depende apenas da forma escolhida para subdividir o intervalo (a,b). Portanto, quando calculamos diferenças em H para diferentes distribuições, dijamos $f(x)$ e $g(x)$, o termo divergente se cancela ao tomarmos a diferença:

$$\Delta H = \lim_{n\to\infty} H(n) = -\int_a^b f(x)\log f(x)\,dx + \int_a^b g(x)\log g(x)\,dx$$

A distribuição uniforme de posições

Consideremos uma partícula confinada em uma caixa unidimensional de tamanho L. Desejamos obter o máximo da função H

$$H = -\int_0^L f(x) \log f(x)\, dx$$

com a condição

$$\int_0^L f(x)\, dx = 1$$

Temos que usar a técnica dos multiplicadores de Lagrange para maximizar H.

$$F = -\int_0^L f(x) \log f(x)\, dx + \lambda\left[\int_0^L f(x)\, dx - 1\right]$$

de forma que $\delta F = 0$

$$\delta F = \int \left[-\log f(x) - 1 + \lambda \right] \delta f(x)\, dx = 0$$

como $\delta f(x)$ é arbitrário,

$$-\log f_{eq}(x) - 1 + \lambda = 0$$

então

$$f_{eq}(x) = e^{\lambda - 1} \quad \text{(escolhendo a base } e\text{)}$$

Pela normalização temos,

$$1 = \int_0^L f_{eq}(x)\,dx = e^{\lambda-1}\int_0^L dx = e^{\lambda-1}L$$

ou

$$e^{\lambda-1} = \frac{1}{L} \Longrightarrow f_{eq}(x) = \frac{1}{L}$$

Portanto,

$$H = -\int_0^L f_{eq}(x)\log f_{eq}(x) = -\frac{1}{L}\log\frac{1}{L}\int_0^L dx = \log L$$

A densidade de probabilidade é uniforme no intervalo de comprimento L. A probabilidade de encontrarmos a partícula entre x e $x+dx$ é:

$$f_{eq}(x)dx = \frac{dx}{L}$$

que não depende de x. Este resultado não é nenhuma surpresa, uma vez que não existe nenhuma posição privilegiada para a partícula estar no interior da caixa.

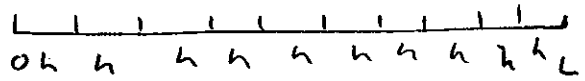
A generalização para 3 dimensões é imediata

$$f_{eq}(x,y,z)\,dx\,dy\,dz = \frac{dx\,dy\,dz}{V}$$

e

$$H = \log V$$

É claro que quanto maior for L, maior será a incerteza na localização da partícula. Vamos denotar a entropia de informação neste caso por $H(L)$. Dividamos agora o intervalo $[0, L]$ em n segmentos de tamanho h.



Usando a propriedade de consistência:

$$H(L) = H\left(\frac{1}{n}, \cdots, \frac{1}{n}\right) + \sum_{i=1}^{n} \frac{1}{n} H(h)$$

$$= \log n + \sum_{i=1}^{n} \frac{1}{n} H(h)$$

$$= \log \frac{L}{h} + \log h = \log L$$

Suponhamos agora que h seja muito pequeno e não estejamos interessados na localização precisa dentro da caixa de comprimento h. Estamos interessados apenas em qual das n caixas a partícula se encontra. Neste caso a entropia de informação se reduz ao caso discreto:

$$H\left(\frac{1}{n}, \cdots, \frac{1}{n}\right) = \log n = \log \frac{L}{h}$$

ou seja, isto é igual a:

$$\Delta H = H(L) - \log h$$

$$= H\left(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n}\right) + \log h - \log h$$

Desta forma, ΔH é uma diferença de entropias e não sofre de divergências. Este resultado é importante no contexto da mecânica estatística de sistemas complexos.

Prova da Forma da Função H

Aqui, vamos considerar que a função H satisfaz os 3 requisitos vistos anteriormente e mostrar como a forma dessa função pode ser obtida.

Temos um experimento com n resultados $A_1, \dots, A_n$, que iremos considerar como sendo mutuamente exclusivos. Vamos agrupa-los em r conjuntos, cada um contendo m_k elementos $(k = 1, 2, \dots, r)$ e $\sum_{k=1}^{r} m_k = n$.

Vamos denotar esses novos eventos $A_1', A_2', \dots A_r'$. Esses eventos são definidos em termos dos eventos originais como:

$$A_1' = \{ A_1 \cup A_2 \cup A_3 \dots \cup A_{m_1} \}$$

$$A_2' = \{ A_{m_1+1} \cup A_{m+2} \cup \dots \cup A_{m_1+m_2} \}$$

$$\vdots$$

$$A_r' = \{ A_{\sum_{k=1}^{r-1} m_{k+1}} \cup \dots \cup A_{n = \sum_{k=1}^{r} m_k} \}$$

Como os eventos originais são considerados mutuamente exclusivos, as probabilidades dos novos eventos serão simplesmente a soma das probabilidades dos eventos originais, então.

$$P_1' = P(A_1') = \sum_{i=1}^{m_1} P_i$$

$$P_2' = P(A_2') = \sum_{i=m+1}^{m_1+m_2} P_i$$

$$\vdots$$

$$P_n' = P(A_n') = \sum_{i=\sum_{k=1}^{n-1} m_{k+1}}^{n} P_i$$

Por conveniência, vamos denotar os eventos incluídos no k-ésimo grupo como $A_1^k\ A_2^k \cdots A_{m_k}^k$, e as suas probabilidades correspondentes como $P_1^k\ P_2^k \cdots P_{m_k}^k$ (o superscrito indica o grupo e o subescrito o elemento desse grupo).

Utilizando esta notação, a propriedade de consistência da função H pode ser escrita como:

$$H(P_1, \cdots P_n) = H(P_1' \cdots P_n') + \sum_{k=1}^{n} P_k'\, H\left(\frac{P_1^k}{P_k'}, \cdots, \frac{P_{m_k}^k}{P_k'}\right)$$

Ou seja, a informação ausente do conjunto original de eventos deve ser igual à informação ausente do novo conjunto de eventos mais a média nos grupos eventos $k = 1, 2, \cdots n$. É importante notar que, $\frac{P_j^k}{P_k'}$ é a probabilidade de ocorrência do j-ésimo evento, considerando-se agora apenas os m_k eventos que compõem o evento k. Ou seja,

$$\sum_{j=1}^{m_k} \frac{P_j^k}{P_k'} = \frac{1}{P_k'} \sum_{j=1}^{m_k} P_j^k = \frac{P_k'}{P_k'} = 1$$

Consideremos agora que as probabilidades do conjunto inicial de eventos sejam todas números racionais, i.e., que elas possam ser escritas como:

$$p_i = \frac{M_i}{\sum_{j=1}^{n} M_j}$$

onde M_i são números inteiros não-negativos. Esta hipótese não é muito forte, uma vez que nós conhecemos as probabilidades dentro de uma precisão ou acurácia, e nós podemos sempre representar isto como um número racional.

Em seguida, vamos "expandir" o nosso conjunto de eventos. Ao invés do nosso conjunto original de n eventos $A_1, A_2, \cdots, A_n$, nós construímos um novo conjunto de eventos no qual o evento A_1 é composto por M_1 elementos de probabilidades iguais a $\frac{1}{M}$, A_2 consiste de M_2 elementos de probabilidades iguais a $\frac{1}{M}$, etc, sendo que

$$M = \sum_{i=1}^{n} M_i$$

Com esta expansão em M eventos, todos com igual probabilidade, nós podemos escrever o requisito de consistência do conjunto de eventos expandido como

$$H\left(\frac{1}{M}, \cdots, \frac{1}{M}\right) = H(p_1, \cdots, p_n) + \sum_{i=1}^{n} p_i H\left(\frac{1/M}{M_i/M}, \cdots, \frac{1/M}{M_i/M}\right)$$

Notemos que na expressão acima, nós temos que o lado esquerdo da equação representa a informação ausente de M eventos equiprováveis. Estes eventos podem ser reagrupados de forma a fornecer os eventos originais.

Vamos agora definir a função,

$$F(M) = H\left(\frac{1}{M}, \cdots, \frac{1}{M}\right),$$

então podemos reescrever acima como

$$F(M) = H(p_1, \cdots, p_n) + \sum_{i=1}^{n} p_i F(M_i)$$

Portanto, se pudermos determinar a forma funcional de $F(M)$, podemos utilizar a equação acima para determinar a forma de $H(p_1, \cdots, p_n)$ para a distribuição $p_1, \cdots, p_n$

A fim de determinarmos a forma de $F(M)$, vamos escolher o caso particular em que todos os M_i são iguais, i.e.,

$$M_i = m \;, \quad \sum_{i=1}^{n} M_i = n\,m = M$$

Neste caso particular as probabilidades são

$$p_i = \frac{M_i}{M} = \frac{m}{M} = \frac{1}{n}$$

Portanto, para este caso particular, temos

$$H(P_1, \cdots, P_n) = H\left(\frac{1}{n}, \cdots, \frac{1}{n}\right) = F(n)$$

e

$$\sum_{i=1}^{n} P_i F(M_i) = \sum_{i=1}^{n} \frac{1}{n} F(m) = F(m)$$

Portanto, pelo que já vimos

$$F(M) = F(n) + F(m)$$

Mas como $M = n \times m$

$$F(n \times m) = F(n) + F(m)$$

Pode-se mostrar facilmente que a única função que possui esta propriedade é a função logaritmo.

Prova:

$$f(xy) = f(x) + f(y)$$

Fazendo $y = 1$

$$f(xy) = f(x) = f(x) + f(1) \Longrightarrow f(1) = 0$$

Seja $y = \frac{x+h}{x}$

Então:

$$f\left(x \cdot \frac{x+h}{x}\right) = f(x) + f\left(\frac{x+h}{x}\right)$$

ou

$$f(x+h) - f(x) = f\left(1 + \frac{h}{x}\right)$$

Dividindo os dois lados da equação por h e multiplicando e dividindo o lado direito por x temos

$$\frac{f(x+h) - f(x)}{h} = \frac{f\left(1 + \frac{h}{x}\right)}{h/x} \cdot \frac{1}{x}$$

tomando o limite $h \to 0$ temos

$$\lim_{h \to 0} \frac{f(x+h) - f(x)}{h} = \frac{1}{x} \lim_{h \to 0} \frac{f\left(1 + \frac{h}{x}\right)}{\frac{h}{x}}$$

então

$$f'(x) = \frac{1}{x} \lim_{h \to 0} \frac{f\left(1 + \frac{h}{x}\right) - f(1)}{\left(1 + \frac{h}{x}\right) - 1} = \frac{1}{x} f'(1)$$

Então,

$$f'(x) = \frac{1}{x}$$

e

$$f(x) = \int f'(x)\,dx = \int \frac{1}{x}\,dx = \ln x + C$$

mas como $f(1) = 0 \Rightarrow C = 0$

Finalmente,

$$f(x) = \ln x$$

Como $\log x = \log_e \ln x$, para qualquer base então

$$f(x) = \log x$$

Retornando à nossa questão original temos

$$F(M) = \log M$$

Desta forma temos,

$$H(p_1, \ldots, p_n) = F(M) - \sum_{i=1}^{n} p_i F(M_i)$$

$$= \log M - \sum_{i=1}^{n} p_i \log M_i$$

$$= \log M \underbrace{\sum_{i=1}^{n} p_i}_{1} - \sum_{i=1}^{n} p_i \log M_i$$

$$= \sum_{i=1}^{n} p_i \log M - \sum_{i=1}^{n} p_i \log M_i$$

$$= \sum_{i=1}^{n} p_i \log \frac{M_i}{M} = \sum_{i=1}^{n} p_i \log p_i$$

Encontramos, portanto, a forma geral da função H:

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

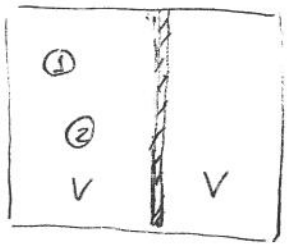
para qualquer distribuição $p_1, \ldots, p_n$.

Exemplo: Gás Ideal

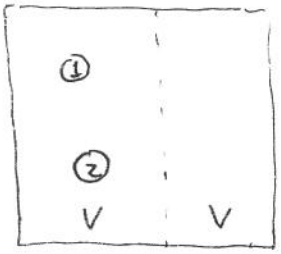
Vamos considerar o caso da expansão livre de um gás ideal que dobra o seu volume no processo. Entendemos um gás ideal como composto por partículas clássicas e distinguíveis que não interagem entre si.

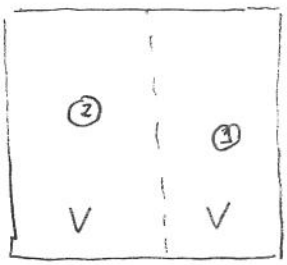
Consideremos inicialmente o caso de um gás composto por 2 partículas.

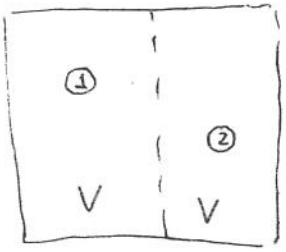
Situação inicial:

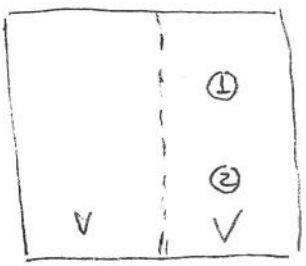


Após a remoção da partição, podemos ter as seguintes situações:





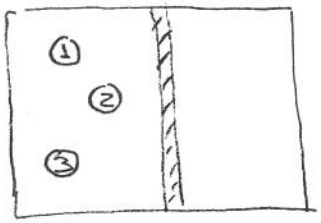




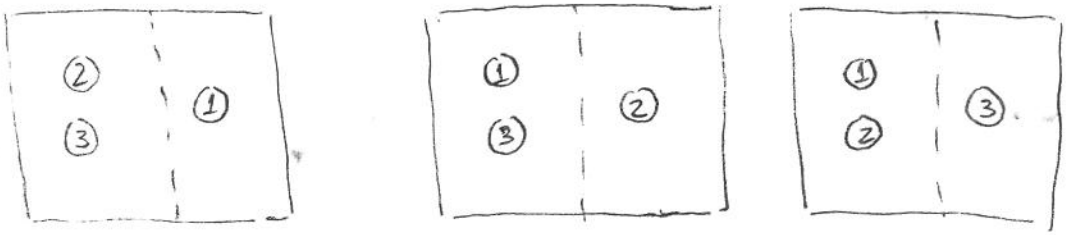
Ou seja, partimos de uma situação em que as partículas se encontram no lado esquerdo do recipiente, havendo apenas uma única possibilidade de localização das partículas e chegamos a uma situação em que há quatro possibilidades para a localização das partículas.

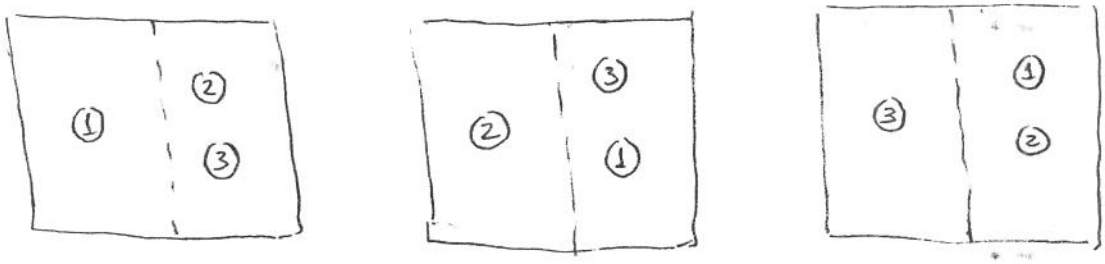
No caso de termos 3 partículas, teremos após a remoção da partição 8 possíveis estados de localização das partículas.

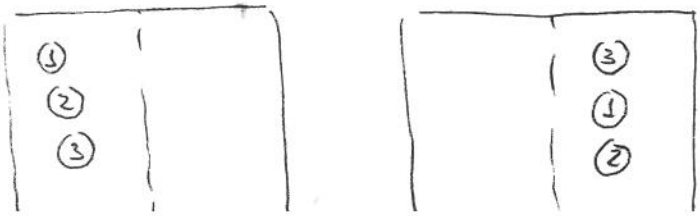
Inicialmente:



Após a expansão:



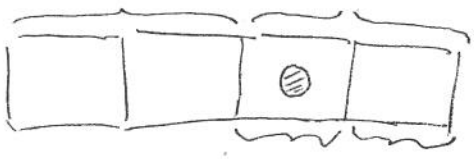




É fácil de generalizar esta multiplicidade de estados de localização após a expansão livre, resultando em 2^N, onde N é o número de partículas.

Portanto, este problema pode ser exatamente mapeado no problema de localizarmos uma moeda em um conjunto de 2^N caixas.

No caso em que temos $N=2$ partículas, o problema se resume a 4 caixas:



Se todas as possibilidades forem equiprováveis, de acordo com a estratégia eficiente, teremos que fazer 2 perguntas binárias

1) A moeda (ou o sistema) se encontra na metade esquerda ou na metade direita do conjunto de 4 caixas. Resposta: metade direita.
2) A moeda (ou o sistema) se encontra na caixa da esquerda ou na caixa da direita. Resposta: caixa da direita.

No caso de 3 partículas, teremos 8 estados. Portanto, teremos que fazer 3 perguntas binárias.

Se tivermos n caixas, o número de perguntas binárias será dado por:

$$H(n) = \log_2 n = \log_2 2^N = N$$

Ou seja, o número de perguntas será igual ao número de partículas.

Como todas as caixas ou todos os estados são equiprováveis, então a probabilidade do gás se encontrar no i-ésimo estado será:

$$P_i = \frac{1}{n}$$

Podemos então escrever no caso geral:

$$H(n) = \frac{n}{n} \log_2 n = - n \frac{1}{n} \log_2 \frac{1}{n}$$

$$H(P_i) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

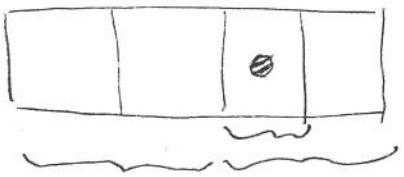
Esta é a expressão obtida por Shannon para quantificar a incerteza associada a uma distribuição de probabilidade.

No caso particular que estamos tratando:

$$H(P_i) = - \log_2 P_i$$

A incerteza ou entropia de informação é medida em bits, ou seja, o número de perguntas binárias que precisamos fazer de acordo com a estratégia eficiente.

Por exemplo, no caso de termos 2 partículas, se considerarmos apenas a pergunta o sistema se encontra na metade esquerda ou na primeira metade dos estados, a sequência de bits 01 significa que o sistema se encontra no terceiro estado, ou seja, o sistema não se encontra na primeira metade do total de estados disponíveis e o sistema se encontra na primeira metade da segunda metade do total de estados



A entropia de informação não precisa necessariamente ser medida em bits, assim como um comprimento pode ser medido em centímetros ou polegadas. Turing, por exemplo, utilizou os logaritmos na base 10 e a entropia de informação era medida em bans. Podemos usar também os logaritmos neperianos (base e), sendo medida em nats.
Desta forma, a relação entre essas medidas é dada por:

$$H_{bit}(n) = \log_2 e \, H_{nat}(n)$$

pois, $\log_2 x = \log_2 e \ln x$

Retornemos ao problema da expansão de gás do ponto de vista termodinâmico.

Pela primeira da termodinâmica podemos escrever:

$$dU = TdS - PdV$$

$$dS = \frac{1}{T} dU + \frac{P}{T} dV$$

A energia interna do gás ideal depende apenas da temperatura, e não do volume. Portanto, na expansão livre do gás ideal T é constante e $dU = 0$.

Assim sendo,

$$dS = \frac{P}{T} dV$$

Mas pela equação de estado do gás ideal

$$\frac{P}{T} = \frac{nR}{V} = \frac{N}{N_0} R \frac{1}{V} = N k_B \frac{1}{V}$$

onde aqui n é o número de moles, N_0 é o número de Avogadro e $k_B = \frac{R}{N_0}$ é a constante de Boltzmann.

Portanto, a variação de entropia termodinâmica do gás será:

$$\Delta S = \int_V^{2V} N k_B \frac{1}{V} dV = N k_B \ln \frac{2V}{V} = N k_B \ln 2$$

ou

$$\Delta S = k_B \ln 2^N$$

No caso de um gás de N partículas, antes da expansão temos apenas um estado possível, as N partículas se encontram na metade esquerda do recipiente, portanto não há incerteza!

$$H(1) = \ln 1 = 0 \quad \text{(medida em nats)}$$

Portanto, a variação na entropia de informação após a expansão será

$$\Delta H = H(n) - H(1) = \ln n = \ln 2^N = N \ln 2$$

Podemos ver então que a variação da entropia termodinâmica e a variação da entropia de informação diferem apenas pelo fator k_B. O fator k_B é o que dá a dimensão de energia/Kelvin da entropia termodinâmica. Enquanto, a entropia de informação é um número puro, pois bits, bans e nats não são dimensões físicas. Na verdade, se a temperatura T fosse medida em unidades de energia, isso não afetaria em nada os resultados da termodinâmica, e a entropia termodinâmica também seria um número puro, pois apenas o produto TS deve ter dimensão de energia.